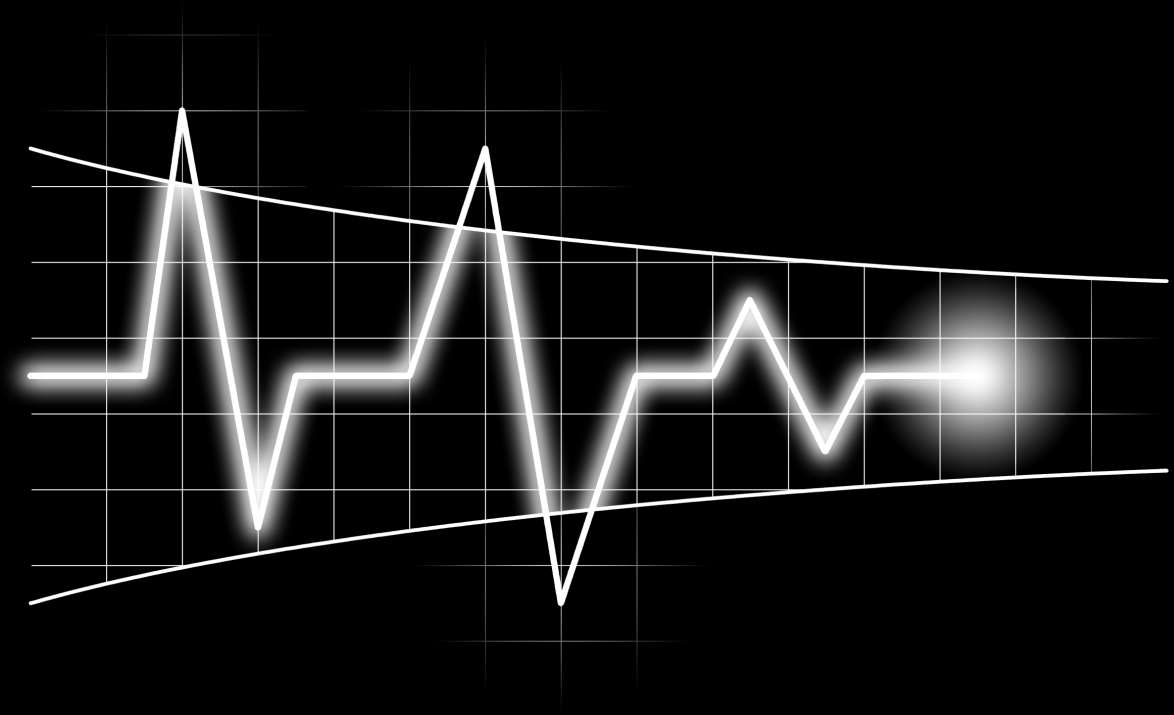
Service-Level Variability and Impatience in Call Centers

Roubos, A.

2012

document version
Publisher's PDF, also known as Version of record

Service-Level Variability and Impatience in Call Centers



Alex Roubos

Service-Level Variability and Impatience in Call Centers

Roubos, Alex, 1985 –
Service-Level Variability and Impatience in Call Centers
ISBN: 978-94-6191-267-1

Printed by Ipskamp Drukkers, The Netherlands.

VRIJE UNIVERSITEIT

Service-Level Variability and Impatience in Call Centers

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. L.M. Bouter,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Exacte Wetenschappen
op maandag 25 juni 2012 om 13.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Alex Roubos

geboren te Amstelveen

promotor: prof.dr. G.M. Koole

Preface

The thesis in front of you is the result of four years of work. Although the content is about mathematical models of call centers, many conclusions are relevant for call center operations management as well. As such, one should not be afraid to continue reading this thesis even if mathematical knowledge is lacking in certain areas.

Let me take this opportunity to explain the artwork on the cover. This illustration depicts an abstract view on the service-level variability. The curve represents a realization of the average service level up to each point in time. A typical observation is that the average service level experiences a high level of variability at the start of the day, where large deviations from the long-run average are likely to occur. At the end of the day, these fluctuations do not occur that often anymore. The upper and lower lines are the boundaries that contain most realizations.

This thesis would not have existed without the continuous support of my parents and the rest of the family. In particular, I would like to thank my brother Dennis for his help and for laying down the footprints for me to follow.

My gratitude also goes toward the people I have had the pleasure of collaborating with throughout the years: René Bekker, Sandjai Bhulai, Oualid Jouini, Ger Koole, and Raik Stolletz. Even though my name is the only one to appear as the author, they all significantly contributed to the work in its present form.

Finally, I would like to thank the members of my thesis committee: Sem Borst, Vijay Mehrotra, Raik Stolletz, and Bert Zwart. Their invaluable comments have helped improve the quality of this thesis.

Alex Roubos
April 2012

Table of Contents

Chapter 1

Introduction

The title of this thesis contains the two elements *service-level variability* and *impatience*. Both elements will be considered in the context of *call centers*.

Service-level variability is related to the fact that most call center models only consider long-run time-average behavior, in which all information about the service level is lost except for the expected service level. This would not be a problem if the variability was low. In all practical situations however, where the service level is considered on a small time interval, this is not the case. Therefore, it is necessary to look at this problem differently.

Another aspect of call center modeling is impatience. Customers that seek contact with a customer service representative often have to wait in a virtual queue before receiving service. Of course, customers are not willing to wait indefinitely: customers are impatient. Basic call center models ignore the presence of impatience. More advanced models are required to deal with impatience and the effect thereof on the performance measures.

1.1 Call Centers

A call center can loosely be defined as a group of agents whose principal business is talking on the telephone to customers. Call centers can be divided into two (overlapping) categories, which are inbound and outbound call centers. In inbound call centers a call is initiated by a customer, whereas in outbound call centers the initiative is with the call center. A further classification is whether the agents possess a single skill or multiple skills. Effectively, this corresponds to whether or not all agents are able to handle all calls. In a single-skill setting there is only one group of agents. In a multi-skill environment there is at least one group of agents

specialized into handling a part of all calls. The scope of this thesis is on inbound and single-skill call centers.

On the one hand, customers have to wait when no agent is available. On the other hand, because calls are initiated by customers, there is also waiting at the side of the agents. The call center management has to balance both types of waiting in order to achieve high agent utilization constrained by a target objective on the waiting time of customers. This is no trivial task because the call center operates under a high level of uncertainty.

The main source of uncertainty is the variability in the number of arrivals over a day. A statistical study of Brown et al. (2005) showed that the arrival process can be described by a nonhomogeneous Poisson process. That is, the arrival rate is approximately piecewise constant for small blocks of time, which are typically 15 or 30 minutes long. Hence, the number of arrivals over a day is Poisson distributed, for which the variance is equal to the expectation. This is a part of the uncertainty that is completely unavoidable. Another part of uncertainty is related to the service process. Brown et al. (2005) claim that the service times are lognormally distributed, although this cannot statistically be validated.

Mathematical models can be constructed to deal with the randomness that is observed in the number of arrivals and in the service times. However, a majority of the current existing models focus only on long-run time-average performance analysis. Models focusing on time-dependent analysis are scarce, and models where performance is aggregated over an interval of finite length are nonexistent. Further sources of variability include differences between agents' speed and quality of answering (agent inhomogeneity), and agent absenteeism. These factors are never taken into account in well-known models. This shows that there still is considerable work to be done in the area of call centers.

1.2 The Basic Call Center Model

The most simple model for a call center is the $M/M/s$ queueing system. Customers arrive to the system according to a Poisson process with arrival rate λ. There are s statistically independent and identical servers available. If a server is free upon the arrival of a customer, then that customer is taken into service. Otherwise, the customer joins a queue with infinite buffer capacity. Service times are exponentially distributed with service rate μ, and customers leave the system after service. When a server completes its service, the first customer in the queue is immediately taken

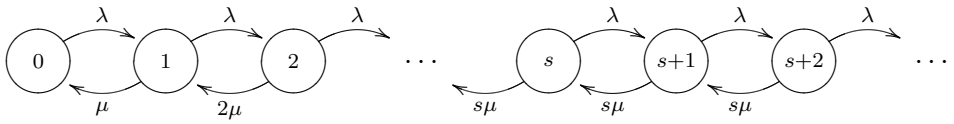


Figure 1.1. Transition diagram of the $M/M/s$ queueing system.

into service. This means that customers are served on a first-come first-served (FCFS) basis. Moreover, customers in the queue wait patiently before service.

The $M/M/s$ queueing system can be analyzed by modeling it as a continuous-time Markov chain (CTMC). Let the state $X(t)$ denote the number of customers present in the system at time t. The stochastic process $\{X(t), t \geq 0\}$ is then a CTMC, with state space $\mathcal{X} = \{0, 1, 2, \dots\} = \mathbb{N}_0$. The transition rates are graphically depicted in the transition diagram in Figure 1.1. Let $a = \lambda/\mu$ be the offered load. Solving the local balance equations for the long-run time-average probability distribution π of the number of customers in the system, gives

$$\pi_i = \begin{cases} \dfrac{a^i}{i!}\pi_0, & 0 \leq i \leq s, \\ \dfrac{a^i}{s!s^{i-s}}\pi_0, & i > s, \end{cases}$$

with

$$\pi_0^{-1} = \sum_{i=0}^{s} \frac{a^i}{i!} + \frac{a^{s+1}}{s!(s-a)}.$$

This stationary distribution π only exists if the system is stable, i.e., on average, per unit of time less work arrives than the servers can handle. This is the case if $\rho = \lambda/(s\mu) < 1$, where ρ is called the offered load per server, or utilization.

A key performance indicator is W_Q, the time an arbitrary customer spends waiting in the queue before service. The expected waiting time and waiting-time distribution are given by

$$\mathbb{E}W_Q = \frac{C(s,a)}{s\mu - \lambda},$$
$$\mathbb{P}(W_Q > \tau) = C(s,a)e^{-(s\mu-\lambda)\tau},$$

with $C(s,a) = \sum_{i=s}^{\infty} \pi_i$. The constant $C(s,a)$ can be interpreted as the probability of delay. These results can be found in many standard books on queueing theory (e.g., Kleinrock 1976). An easy way to compute the probability of delay is by relating it to the probability of blocking in the $M/M/s/s$ queue, where customers are blocked if upon arrival all servers are occupied. Cooper (1981) gives the following relation

$$C(s,a) = \frac{sB(s,a)}{s - a(1 - B(s,a))},$$
$$B(s,a) = \mathbb{P}(N = s)/\mathbb{P}(N \leq s),$$

where $N \sim \text{Poisson}(a)$, for $s > a$.

The service level (SL) is defined as the fraction of customers with a waiting time in the queue no longer than τ time units. The parameter τ is also called the acceptable waiting time (AWT). On the long run, in a stationary situation, the service level can be interpreted as the probability that the waiting time in the queue of an arbitrary customer is less than or equal to τ. That is,

$$\text{SL} = \mathbb{P}(W_Q \leq \tau).$$

This is called the Erlang C formula. The Erlang C formula is frequently used to seek the minimum number of agents such that the service level is above a certain threshold. The industry standard is 80/20, which means that at least 80% of the customers should wait no longer than 20 seconds, i.e., $\mathbb{P}(W_Q \leq 20 \text{ seconds}) \geq 0.8$.

1.3 Call Center Model with Abandonments

The basic call center model can be extended to a model that includes impatient customers. When a customer has to wait in the queue, instead of waiting patiently before service, that customer has a stochastic limit on the waiting time, i.e., his or her patience. If service has not started before this limit is reached, the customer abandons the queue and leaves the system unserved. Patience is modeled by the random variable T that is exponentially distributed with parameter γ. This model is denoted by the $M/M/s + M$ queueing system, and also goes by the name of the Erlang A model.

Analysis of the $M/M/s + M$ queueing system is in line with the $M/M/s$ queueing system. Let $X(t)$ denote the number of customers in the system at time

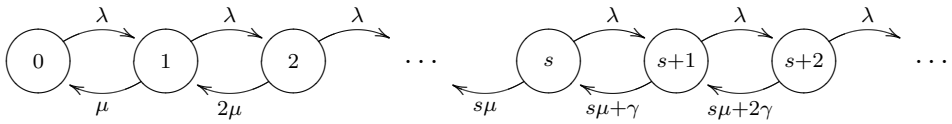


Figure 1.2. Transition diagram of the $M/M/s+M$ queueing system.

t. Then, the stochastic process $\{X(t), t \geq 0\}$ is a CTMC with state space $\mathcal{X} = \mathbb{N}_0$. The transition diagram is displayed in Figure 1.2. The stationary distribution, which follows from solving the local balance equations, is given by

$$\pi_i = \begin{cases} \dfrac{\lambda^i}{\mu^i i!}\pi_0, & 0 \leq i \leq s, \\ \dfrac{\lambda^i}{\mu^s s! \prod_{j=1}^{i-s}(s\mu + j\gamma)}\pi_0, & i > s, \end{cases}$$

with

$$\pi_0^{-1} = \sum_{i=0}^{s} \frac{\lambda^i}{\mu^i i!} + \sum_{i=s+1}^{\infty} \frac{\lambda^i}{\mu^s s! \prod_{j=1}^{i-s}(s\mu + j\gamma)}.$$

Although the stationary distribution has an analytical solution, it is not very tractable. It is more convenient to calculate the stationary distribution numerically. For that purpose, let $\tilde{\pi}$ be the intermediate solution that satisfies the local balance equations, and does not yet satisfy the normalizing condition. Let $\tilde{\pi}_s = 1$, then, for $i = s, s-1, \ldots, 1$,

$$\tilde{\pi}_{i-1} = \frac{i\mu}{\lambda}\tilde{\pi}_i,$$

and, for $i = s, s+1, \ldots,$

$$\tilde{\pi}_{i+1} = \frac{\lambda}{s\mu + (i-s+1)\gamma}\tilde{\pi}_i.$$

The stationary distribution can then be obtained by normalizing, i.e.,

$$\pi_i = \frac{\tilde{\pi}_i}{\sum_{j=0}^{\infty}\tilde{\pi}_j}.$$

To implement this, one has to circumvent the issue of the infinite series. The state space can always be truncated to a sufficiently large constant K, such that $\sum_{i=0}^{K} \pi_i > 1 - \epsilon$, for any $\epsilon > 0$.

In contrast with the basic call center model, this model with impatient customers is always stable. Customers will simply abandon when they believe that they are not served quickly enough. Let A be the event that an arbitrary customer will abandon. It holds that

$$\mathbb{P}(A) = \sum_{i=s+1}^{\infty} \frac{(i-s)\gamma\pi_i}{\lambda}.$$

This result follows from the interpretation of the transition rates that, given that the system is in state $i > s$ with probability π_i, there are $(i-s)\gamma$ abandonments per unit of time. Combining this with the fact that there are on average λ arrivals per unit of time, the abandonment probability follows directly.

The notion of the waiting time is complicated in a queueing system with abandonments. The random variable W_Q denotes the actual waiting time in the queue, incurred by an arbitrary arriving customer. The end of the waiting time is initiated by either the start of service, or abandonment. On the other hand, it is mathematically more appealing to use the virtual waiting time V_Q. The virtual waiting time is defined as the waiting time in the queue incurred by an infinitely patient customer. The relation between the actual waiting time, the virtual waiting time, and the patience is captured by $W_Q = \min\{V_Q, T\}$.

The expected actual waiting time can be obtained from the stationary distribution π. Let L_Q denote the long-run time-average number of customers in the queue. Then,

$$\mathbb{E}L_Q = \sum_{i=s+1}^{\infty} (i-s)\pi_i.$$

Using Little's law (Little 1961), the expected waiting time is

$$\mathbb{E}W_Q = \frac{\mathbb{E}L_Q}{\lambda}.$$

The waiting-time distribution can also be determined. By the law of total probabil-

ity, conditioning on the state of the system seen by an arriving customer,

$$\begin{aligned}\mathbb{P}(W_Q > \tau) &= \lim_{t\to\infty} \sum_{i=0}^{\infty} \mathbb{P}(W_Q > \tau \,|\, X(t) = i)\mathbb{P}(X(t) = i) \\ &= \sum_{i=0}^{\infty} \mathbb{P}(W_Q > \tau \,|\, L_Q = i)\pi_{s+i}.\end{aligned}$$

Finally, using a result of Riordan (1962), which was later refined in Deslauriers et al. (2007), for $i \geq 0$,

$$\mathbb{P}(W_Q > \tau \,|\, L_Q = i) = e^{-(s\mu+\gamma)\tau} \sum_{j=0}^{i} \frac{\phi_j(1-e^{-\gamma\tau})^j}{j!},$$

where $\phi = s\mu/\gamma$, $\phi_0 = 1$, and $\phi_j = \phi(\phi+1)\cdots(\phi+j-1)$ for $j \geq 1$. A similar result holds for the virtual waiting-time distribution. Dividing by the probability that an arriving customer will not abandon before τ, i.e., $e^{-\gamma\tau}$, yields

$$\mathbb{P}(V_Q > \tau \,|\, L_Q = i) = e^{-s\mu\tau} \sum_{j=0}^{i} \frac{\phi_j(1-e^{-\gamma\tau})^j}{j!}.$$

The virtual waiting time will be more elaborately discussed in a more general setting next.

1.3.1 General Patience Distribution

The $M/M/s+G$ queueing system is a model where the patience is not necessarily assumed to be exponentially distributed. Instead, the patience has a general distribution with cumulative distribution function $G(x)$, $x \geq 0$. Based on the work of Baccelli and Hebuterne (1981), Zeltyn and Mandelbaum (2005) define the following building blocks for performance analysis. Let $\bar{G}(x) = 1 - G(x)$. Define

$$\begin{aligned}H(x) &= \int_0^x \bar{G}(u)\mathrm{d}u, \\ J(t) &= \int_t^\infty e^{\lambda H(x) - s\mu x}\mathrm{d}x, \\ J_1(t) &= \int_t^\infty x e^{\lambda H(x) - s\mu x}\mathrm{d}x, \\ J_H(t) &= \int_t^\infty H(x) e^{\lambda H(x) - s\mu x}\mathrm{d}x.\end{aligned}$$

Let $J = J(0)$, $J_1 = J_1(0)$, $J_H = J_H(0)$, and define

$$\mathcal{E} = \frac{\sum_{i=0}^{s-1} \frac{(\lambda/\mu)^i}{i!}}{\frac{(\lambda/\mu)^{s-1}}{(s-1)!}} = B(s-1,a)^{-1}.$$

For simple distribution functions $G(x)$, e.g., any combination of exponential functions, the function $H(x)$ can be determined in closed form. However, there is no hope that one can find a closed-form expression for $J(t)$. Hence, it is required to use numerical techniques in order to evaluate the system's performance.

Expressed in these building blocks, the probability density function of the virtual waiting time V_Q is given by, for $x > 0$,

$$v(x) = \frac{\lambda e^{\lambda H(x) - s\mu x}}{\mathcal{E} + \lambda J},$$

with a mass at the origin with value

$$\mathbb{P}(V_Q = 0) = \frac{\mathcal{E}}{\mathcal{E} + \lambda J}.$$

Zeltyn and Mandelbaum (2005) derive a number of performance measures, including

$$\begin{aligned}
\mathbb{P}(A) &= \frac{1 + (\lambda - s\mu)J}{\mathcal{E} + \lambda J}, \\
\mathbb{E}V_Q &= \frac{\lambda J_1}{\mathcal{E} + \lambda J}, \\
\mathbb{E}W_Q &= \frac{\lambda J_H}{\mathcal{E} + \lambda J}, \\
\mathbb{P}(V_Q > \tau) &= \frac{\lambda J(\tau)}{\mathcal{E} + \lambda J}, \\
\mathbb{P}(W_Q > \tau) &= \frac{\lambda \bar{G}(\tau) J(\tau)}{\mathcal{E} + \lambda J}.
\end{aligned}$$

With abandonments, it is not clear anymore which service-level measure should be used. In fact, there are multiple definitions used in practice. This will be more thoroughly discussed in Chapter 5.

1.4 Structure of the Thesis

The remainder of this thesis is organized as follows.

- In Chapter 2 the observation of variability in the service level is first described and then quantified. In practice, service levels are reported over periods of finite length that are usually no longer than 24 hours. The variability is nonnegligible in such small periods. Staffing decisions should therefore not only be based on the expected service level, but should also take this variability into account. The service-level distribution is quantified by means of an approximate method based on simulations. This distribution is used for a service-level variability-controlled staffing approach to circumvent the shortcomings of the traditional staffing based on the expected service level. Chapter 2 is based on Roubos et al. (2012b).

- While in Chapter 2 the service level is only passively controlled, in Chapter 3 an active control policy is applied. In light of long-term planning, permanent agents are scheduled. Using the idea of flexible agents, these staffing levels can be changed during the day to reach a soft constraint on the service level at the end of the day. This problem is modeled as a Markov decision process, where decisions are based on the realized service level so far. The optimal policy provides a good balance between staffing costs and the penalty probability for not reaching the service level. Chapter 3 is based on Roubos et al. (2011).

- In Chapter 4 the distribution of the service level is considered again. This distribution is approximately quantified in Chapter 2, and also used in Chapter 3 to determine the transition probabilities in the Markov decision process. The customer-average service level is related to the time-average proportion of the time that the virtual waiting-time process is at or below the acceptable waiting time. An exact analysis is presented for the double Laplace-Stieltjes transform of the time below the acceptable waiting time in an interval of finite length. Furthermore, interesting properties of the service-level distribution are derived. Chapter 4 is based on Roubos et al. (2012a).

- Chapter 5 shifts the focus to impatience customers. In models that ignore impatience, it is clear what is meant by the phrase service level. When taking

abandonments into account however, the service level is not unambiguously defined anymore. This chapter studies a number of different service-level definitions, including all those used in practice. Based on data from different call centers, two new models are introduced. The first model is a slight extension to the Erlang A model, in which customers are allowed to balk upon arrival. In the second model, the patience is modeled by the hyperexponential distribution. Both models are shown to fit reality very well. Through numerical analysis, the different models and the different service-level definitions are compared. Chapter 5 is based on Jouini et al. (2011b)

- Chapter 6 deals with multi-server queues with multiple customer classes with impatience. Customers are grouped by their priority, and high-priority customers get nonpreemptive priority over the other type. Besides the FCFS service discipline, the last-come first-served (LCFS) service discipline is also studied. Focus is given to determine performance measures related to queueing delays. This chapter closes with numerical experiments to gain some interesting management insights. Chapter 6 is based on Jouini and Roubos (2011).

Chapter 2

Service-Level Variability of Inbound Call Centers

In practice, call center service levels are reported over periods of finite length that are usually no longer than 24 hours. In such small periods the service level has a large variability. It is therefore not sufficient to base staffing decisions only on the expected service level. In this chapter we consider the classical $M/M/s$ queueing model that is often used in call centers. We develop accurate approximations for the service-level distribution based on extensive simulations. This distribution is used for a service-level variability-controlled staffing approach to circumvent the shortcomings of the traditional staffing based on the expected service level.

2.1 Introduction

The hierarchical planning in call centers is usually divided into forecasting, requirements planning for short intervals, and staff scheduling (see Gans et al. 2003). For the requirements planning, stationary queueing models are used to determine the minimum number of agents to fulfill a specific performance measure. In call centers, the Erlang C model is often used to provide an estimate for the fraction of calls that wait no more than Z seconds. This service-level estimate Y can be interpreted as the long-run fraction of calls that wait no more than Z seconds. However, in call centers we are never interested in the long run: service-level realizations are considered at 30-minute intervals, and sometimes aggregated over full days, but seldom over longer periods (see, e.g., Stolletz 2003). The service-level target that Y% of the calls are answered within Z seconds is commonly expressed in the form Y/Z. The goal of call center managers is often to meet an aggregated Y/Z service level for a high fraction X of periods.

Service levels fluctuate. One of the reasons for service-level deviations is that

call centers operate in a highly volatile environment, with possibly erroneous forecasts, staffing levels that are not as planned, etc. Even if all parameters are correct, the realized service level will still deviate from the service-level prediction, because of the intrinsic randomness in the call center environment. Simulations show that this difference can be considerable, for example, 5% over a whole day is not exceptional (see Section 2.3). Managers are aware that the actual service level can differ from the expected service level. However, they do not realize the impact of the randomness on the amount of fluctuations. It is our personal experience that managers are surprised to learn this and are willing to consider new solutions, such as the one we propose.

Call center managers deal with fluctuations by *traffic management*, the activity that consists of rescheduling the workforce on short notice to obtain the required service level (see, e.g., Mehrotra et al. 2010). A higher than necessary service level is generally not a problem, but managers might be penalized for failing to meet the target in too many periods. To this end, some managers deliberately opt for a higher expected service level $\bar{Y} > Y$ or a lower target time $\bar{Z} < Z$ to meet the original target Y/Z with higher likelihood. Such behavior is also observed in inventory management (Thomas 2005) and other fields. Both approaches are based on the experience of the call center manager, because the influence of $\bar{Y}$ and $\bar{Z}$ on the probability X to reach the target Y/Z is not yet described in the literature.

Costs play a crucial role in our analysis. For example, when staffing according to the expected value of the service level, the target service level may only be met 50% of the time intervals (see Section 2.4). However, one additional staffed agent can already improve this probability to 80%. Is it better to risk not reaching the target service level 50% of the time, or to schedule one additional agent and accept a risk of 20%? To make this trade-off, we have to quantify both the costs of staffing and the costs for not reaching the target service level. Finding the optimal trade-off then becomes equivalent to minimizing total costs. Related to this is the work of Baron and Milner (2009), where approximations are constructed for the expected penalties for failing to meet the target service level for impatient customers.

In call center planning, there are a number of challenging problems related to time-varying arrival rates. For forecasting problems with time-varying rates, we refer to Akşin et al. (2007) and Steckley et al. (2009). The stationary independent period-by-period (SIPP) approach, and variants of it, are widely used for time-dependent requirements planning (staffing) in call centers (see Green et al. 2001, 2003). Ingolfsson et al. (2007) and Stolletz (2008) review these and other evaluation

methods for time-dependent systems and compare them in numerical experiments. In all these methods there is no distinction between the staffing period and the aggregation interval for performance measurement.

The contribution of this chapter is twofold. First, we analyze the variability of the service level as a function of the length of the aggregation interval. For such a finite-length interval, the actual service level is a random variable, and the service-level estimate given by the Erlang C formula is the *expected* service level. We give a closed-form approximation for the complete distribution of the service level and validate it extensively. Second, in contrast to decisions about staffing levels that are based on the expected service level, we propose a new approach for variability-controlled staffing. The approximated distribution of the service level is used to set the staffing level to meet the service level Y/Z with a targeted probability X. We integrate this variability-controlled staffing approach in the traditional SIPP approach for time-dependent rates. With this method the staffing period and the aggregation interval could be different, which is important for highly volatile rates in call centers.

Related to our first contribution is the work of Steckley et al. (2009), who provide an analysis to compute the service-level distribution for the special case $Z = 0$ only. Their approach works if, upon a customer arrival, it can be determined from the state of the system whether that customer will receive service on or before Z. In case $Z = 0$, a customer will receive satisfactory service if at least one server is available. Therefore, the state can be chosen as the number of customers in the system. Their approach cannot be generalized to $Z > 0$.

The remainder of this chapter is organized as follows. We start in Section 2.2 with the model description, where the basic notation and definitions are introduced for the queueing model under consideration. Section 2.3 deals with the approximations that are based on numerical experiments. Several performance evaluations are presented as well. The approximations of Section 2.3 are used in Section 2.4, where we present a new way to do staffing calculations. We do this in such a way that we have desired control over the variability. In Section 2.5 we show how our staffing approach could be used to address the issue of nonhomogeneous systems. Finally, conclusions and directions for further research are given in Section 2.6.

2.2 Model Description

We model a call center by the $M/M/s$ queueing system, i.e., we consider the basic call center model described in Chapter 1. With arrival rate λ, service rate μ, and s servers, we restate the Erlang C formula:

$$\mathbb{P}(W_Q \leq \tau) = 1 - C(s, a)e^{-(s\mu-\lambda)\tau}. \tag{2.1}$$

We will denote $\mathbb{P}(W_Q \leq \tau)$ by the expected service level $\mathbb{E}$SL. The expected service level depends on λ, μ, s, and τ. Traditionally, service-level objectives have been denoted as Y/Z, which means that at least Y% of the customers have to wait less than or equal to Z seconds. Both τ and Z can be used to denote the acceptable waiting time, although the unit of Z is seconds and the time unit of τ can be chosen arbitrarily. There is a difference between $\mathbb{E}$SL and Y: Y is used to denote the target service level, i.e., the minimum required service level; $\mathbb{E}$SL is the service level that is expected to be obtained given all parameters. Although the steady-state performance measure Y/Z will be met in the long run, we are interested in the service level aggregated over intervals of finite length t. The realized average service level could be lower or higher than the expected one. The distribution of the realized average service level strongly depends on the length t.

Throughout this chapter, we assess the accuracy of the approximations and our staffing approach on several examples. We mainly consider two call centers modeled by the $M/M/s$ queueing system, with parameters that could be found in practice. These systems are defined as follows.

Large system $\lambda = 40$, $\mu = 0.2$, and $s = 210$.

Small system $\lambda = 3$, $\mu = 0.2$, and $s = 19$.

Unless specified otherwise, the time scale is expressed in minutes, and we take the acceptable waiting time equal to $\tau = 1/3$. This means that the expected service level is 80.7% for the large system and 81.3% for the small system.

2.3 Numerical Approximations

To demonstrate the effect of the aggregation length t on the service-level distribution, we have performed straightforward simulations of the large system. The results are shown in Figure 2.1. The simulations are performed 10,000 times, each

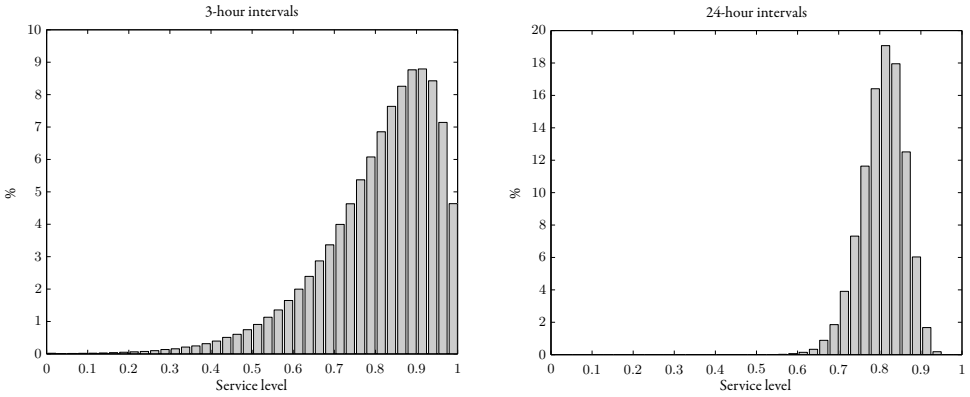


Figure 2.1. Histograms of the service level aggregated over 3-hour intervals on the left and aggregated over 24-hour intervals on the right.

starting after a warm-up period of 24 hours (such that the transient effects of starting from an empty system are gone) and continuing for 3 hours and 24 hours, respectively. After each run, one realization of the service level is obtained. The histograms depict what percentage of the runs fall into each of the bins. In both cases, the average service level is 80.7%, which is equal to the outcome of the Erlang C formula. Consider the complete distribution of the service level. The shape of the distribution depends on the level of aggregation. For a short aggregation length (e.g., 3 hours), the distribution is asymmetric and has a large variability, whereas for a longer aggregation length (e.g., 24 hours), the variability decreases and the distribution looks more like a normal distribution. This can be explained by the central limit theorem (see Baron and Milner 2009, Corollary 2). What is remarkable is that the variability, even when aggregated over the whole day, is still huge: 35% of the realizations deviate more than 5% from the average in this example (i.e., they have a service level outside [75.7%, 85.7%]).

To account for the significant variability of the service level in intervals of finite length, staffing decisions should not only be made on the basis of the expected service level, but should also reflect both the variability that is inherent in service levels and the level of desired confidence in achieving the service-level objective. To be able to do this, we need to quantify this variability. In this section we show that we can approximate the distribution of the service level by the normal distribution. In the normal distribution the variability is characterized by the standard deviation. To this end, we develop an approximation for the standard deviation.

2.3.1 Standard Deviation Approximation

In the limit $t \to \infty$, the service-level distribution approaches the normal distribution. It is intuitively clear (and can also be observed from Figure 2.1) that the standard deviation goes to zero in this limit. On the other hand, the standard deviation is positive for finite t. Furthermore, if t is large enough, the service-level distribution cannot be distinguished from the normal distribution, according to statistical tests for normality (see Subsection 2.3.2 for a description of such a test). As a first step, we therefore consider large t and express the estimate $\hat{\sigma}$ of the unknown standard deviation σ in the system parameters λ, μ, s, τ, and t. We denote that $\hat{\sigma}$ is a function of these parameters by $\hat{\sigma}(\cdot)$. As a next step, we show the results of this approximation for shorter intervals.

The central limit theorem can be used to derive the functional form of σ. The central limit theorem states that the distribution of the average of n independent and identically distributed random variables, each having mean $\mathbb{E}\mathrm{SL}$ and standard deviation ς, converges to the normal distribution with mean $\mathbb{E}\mathrm{SL}$ and standard deviation $\sigma = \varsigma/\sqrt{n}$. Baron and Milner (2009, Corollary 2) prove that the central limit theorem also holds for a stochastic number of random variables. The contributions of the individual customers to the service level are not independent. However, the contributions of renewal cycles are independent.

Consider a renewal process with the epochs at which an arriving customer initiates a busy period as renewal moments. The time between consecutive renewal moments consists of a busy period B and an idle period I, so that the mean time between renewals is $\mathbb{E}B + \mathbb{E}I$. Then, by Asmussen (2003, Proposition 1.4) in the interval of length t, the number of renewal cycles converges to $n = t/(\mathbb{E}B + \mathbb{E}I)$ as $t \to \infty$.

Result for $\tau = 0$

For $\tau = 0$ it is possible to derive the standard deviation ς of the service level in a renewal cycle. In this case, only the customers that arrive during the period in which at least one server is idle, i.e., the idle period, are successfully served. Daley and Servi (1998) give the mean and variance for the number of arrivals in a busy

period, N_B, and in an idle period, N_I. They are

$$\mathbb{E}N_B = \frac{1}{1-\rho}, \qquad \operatorname{var} N_B = \frac{\rho(1+\rho)}{(1-\rho)^3},$$
$$\mathbb{E}N_I = \frac{P_{s-1}}{\pi_{s-1}}, \qquad \operatorname{var} N_I = 2\sum_{i=1}^{s-1} \frac{P_i P_{i-1}}{\pi_i \pi_{s-1}} + \frac{P_{s-1}}{\pi_{s-1}} - \left(\frac{P_{s-1}}{\pi_{s-1}}\right)^2,$$

where π is the steady-state distribution of the number of customers in the system and $P_i = \sum_{j=0}^{i} \pi_j$. The service level is then given by $N_I/(N_I + N_B - 1)$. The -1 comes from the fact that the arrival that initiates the busy period is included in both periods. The expected value of the service level follows immediately from the renewal process and is given by

$$\mathbb{E}\mathrm{SL} = \frac{P_{s-1}/\pi_{s-1}}{P_{s-1}/\pi_{s-1} + \rho/(1-\rho)},$$

which is also equal to the outcome of the Erlang C formula. The variance of the service level in a renewal cycle can be obtained from the multivariate delta method (Casella and Berger 2002, Subsection 5.5.4), i.e., a Taylor series expansion. Using the most important terms in the series expansion, the variance simplifies to

$$\varsigma^2 \approx \frac{\operatorname{var} N_I}{(\mathbb{E}N_I + \mathbb{E}N_B - 1)^2} - 2\frac{\mathbb{E}N_I \operatorname{var} N_I}{(\mathbb{E}N_I + \mathbb{E}N_B - 1)^3} + \frac{(\mathbb{E}N_I)^2(\operatorname{var} N_I + \operatorname{var} N_B)}{(\mathbb{E}N_I + \mathbb{E}N_B - 1)^4}.$$

Finally, the mean length of the renewal cycle equals $(\mathbb{E}N_I + \mathbb{E}N_B - 1)/\lambda$, and hence

$$n = \frac{t\lambda}{\mathbb{E}N_I + \mathbb{E}N_B - 1},$$

as $t \to \infty$. The standard deviation is then approximately given by $\sigma = \varsigma/\sqrt{n}$.

A special case is the $M/M/1$ queue, for which these expressions can be simplified to $\mathbb{E}\mathrm{SL} = 1-\rho$, $\varsigma^2 \approx \rho(1+\rho)(1-\rho)$, $n = t\lambda(1-\rho)$ and $\sigma^2 \approx (1+\rho)/(\mu t)$.

In Steckley et al. (2009) an analysis is provided to approximate the standard deviation in case $\tau = 0$. That approximation has to be obtained by solving multiple sets of equations. We have extended their results by providing a closed-form solution. Both methods give exactly the same standard deviation. This follows from

an analytical comparison in case $s = 1$ and from a numerical comparison in case $s > 1$.

Although this standard deviation approximation for $\tau = 0$ has been analytically derived, numerical results show that it is not accurate for a high utilization in an interval of finite length. For example, if ρ goes to one in the $M/M/1$ queue, the standard deviation goes to $\sqrt{2/(\mu t)}$, which is nonzero. However, one would expect that the standard deviation goes to zero, because there is almost no variability when the expected service level goes to zero. Differences are clearly noticeable for $\rho > 0.5$ for the $M/M/1$ queue. The accuracy increases for systems with more agents. For instance, a system with $s = 10$ has a perfect accuracy for $\rho < 0.9$. That the accuracy decreases at high utilization can be attributed to the application of the central limit theorem. The length of a renewal cycle increases as the utilization increases, and therefore there are fewer independent renewal cycles in a fixed-length interval. Because this approach can only be applied to systems with $\tau = 0$, we take the following alternative approach to approximate the standard deviation for $\tau > 0$.

Method for $\tau > 0$

The method consists of generating the "real" standard deviation σ by means of simulations for different parameter combinations. We then try to find an approximation $\hat{\sigma}$, such that the approximation is very accurate on all generated instances. The parameter combinations used in the simulations are obtained by the following steps.

Step 1 We varied the target service level from the set {0.25, 0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95} and the acceptable waiting time τ from the set {1/6, 1/3, 1/2, 1, 2}.

Step 2 We varied the offered load ρ within the interval $[0.5, 1)$ in step sizes of 0.001, and we fixed μ equal to 0.25.

Step 3 A unique combination of the pair (λ, s) exists for given values of ρ and μ such that the expected value of the service level is as close as possible to the Y/Z service level chosen in Step 1. After this step, the s remains fixed.

Step 4 Because of the integrality constraint of s, however, the expected service level might not be close enough to the target. For given values of ρ and s, we generally can get arbitrarily close by changing μ and hence λ. To be

	λ	μ	s	τ	t
Lower bound	0.1	0.2	1	1/6	6,000
Upper bound	200	2	750	2	6,000

Table 2.1. Bounds of the parameter combinations used for approximating σ.

precise, we increase μ by a step size until the expected service level is greater than the target. In this case, we halve the step size and start decreasing μ until the expected service level is lower than the target. We continue until we reach the Y/Z service level within the desired accuracy of 0.001. The only exception is that for very lightly loaded systems the s computed in Step 3 might already be too high to ever reach the target. We ignored these instances.

Table 2.1 lists the bounds of the parameter combinations that we have obtained using this scheme. Note that we have a value of $t = 6{,}000$ for the aggregation interval, which is large enough for the normal distribution to be justified. In total we have performed well over 20,000 different simulations. Each simulation is independently executed 1,000 times, out of which one simulated standard deviation of the service level is obtained. Again, the warm-up period is 24 hours.

In this way we have the standard deviation for a wide range of parameter combinations. The goal is to construct a function $\hat{\sigma}$ that can very accurately fit the data.

Result for $\tau > 0$

Motivated by the simulation results for a fixed service level and acceptable waiting time, we deduce the following simple functional form to describe the data:

$$\hat{\sigma}(\lambda, \mu, s, \tau, t) = \frac{\alpha(\mathbb{E}\mathrm{SL}, \tau)}{\sqrt{s\mu}(1-\rho)\sqrt{t}}, \tag{2.2}$$

where α is a parameter that depends on the system parameters only through the expected service level and the acceptable waiting time. To approximate α, we impose the functional form given by $\alpha(\mathbb{E}\mathrm{SL}, \tau) = (1 - \mathbb{E}\mathrm{SL})^{a_1+a_2\tau} \times (\mathbb{E}\mathrm{SL})^{b_1+b_2\tau} \times (c_1 + c_2\tau)$. This specific form is motivated by our observations in the data and the requirement that the standard deviation is zero in the case when the expected

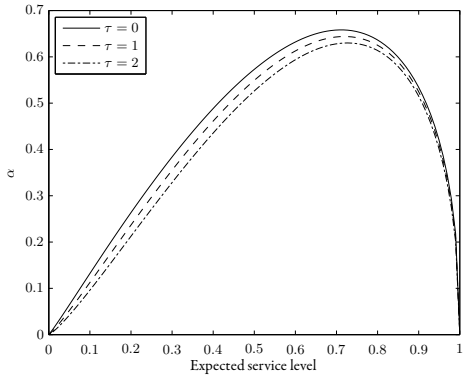


Figure 2.2. Plot of the function α dependent on the expected service level for different values of τ (in minutes).

service level is either zero or one. The constants are determined by the least-squares regression over all experiments. In the end, α is given by

$$\begin{aligned}\alpha(\mathbb{E}\mathrm{SL}, \tau) = (1 - \mathbb{E}\mathrm{SL})^{0.4348+0.0132\tau} \times (\mathbb{E}\mathrm{SL})^{1.0708+0.0776\tau} \\ \times (1.6271 + 0.0339\tau).\end{aligned} \tag{2.3}$$

The corresponding mean squared error is then $4.4 \cdot 10^{-6}$. In addition, the mean absolute percentage error is only 3.4%, despite the divisions by very small numbers. The value of the coefficient of determination, defined by $R^2 = 1 - \sum_i (\sigma_i - \hat{\sigma}_i)^2 / \sum_i (\sigma_i - \bar{\sigma})^2$, is 0.98.

Figure 2.2 shows how the value of α depends on the expected service level and the acceptable waiting time. If the expected service level is close to its bounds of zero or one, i.e., a really bad or an excellent customer service, the value of the parameter α is close to zero. Also, for increasing values of the acceptable waiting time, the parameter α decreases.

It is possible to apply this approximation to the case $\tau = 0$ as well. Then, we observe an increased accuracy at high utilizations, compared to the analytical approximation for $\tau = 0$ that loses accuracy at high utilizations. In that sense, this approximation is an important addition to the analytical one, even for $\tau = 0$.

Validation

To validate Equation (2.2), we simulated 200 new instances that are shown in Figure 2.3. This figure shows the simulated and approximated standard deviation

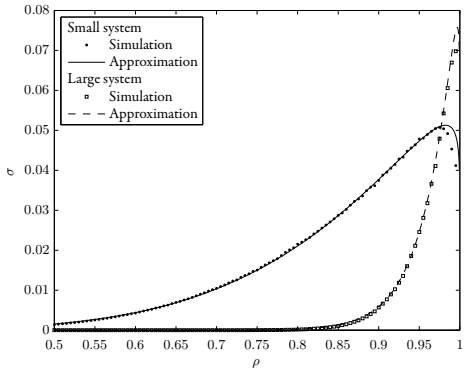


Figure 2.3. Comparison of the simulated and approximated standard deviation for the small and large systems.

σ for the small and large systems, dependent on the utilization ρ. The arrival rate λ is changed from the base examples such that the specified ρ is obtained. This plot shows that the standard deviation is well approximated for a broad range of ρ. Only in case of an unrealistically high utilization is the standard deviation overestimated. In these cases ($\rho > 0.98$), the expected service level is way below 50%, so there are more important concerns other than a well-approximated standard deviation. The standard deviations increase if ρ increases up to a very high utilization before it starts to diminish.

Next we show the generality of the approximations. We consider parameter combinations chosen at random uniformly between the lower and upper bounds displayed in Table 2.1. The interval length t is chosen from $[600, 6{,}000]$ instead to allow other large intervals as well. Randomly chosen parameter combinations can result in unstable systems. Therefore, we only considered stable systems. In addition, we considered systems with an expected service level less than one only. Otherwise, the standard deviation will be zero because there is no variability. After 500 randomly selected instances, we get that the mean value of the simulated standard deviation is $1.1 \cdot 10^{-3}$. Moreover, we obtain a mean absolute error of $6.3 \cdot 10^{-5}$, and the maximum absolute error is $3.0 \cdot 10^{-3}$. The absolute percentage error corresponding to this maximum is only 1.9%. We see that under all circumstances the accuracy of the approximation is very good. For parameter values outside the wide range of values in Table 2.1, our approximation has yet to be validated.

Shorter Intervals

So far, we have considered large values of the interval length t. We have developed an approximation for the standard deviation of the service level, and we have shown that it has excellent accuracy in these cases. Moreover, the distribution of the service level is indistinguishable from the normal distribution.

In shorter intervals the distribution will be different from the normal distribution (see, e.g., Figure 2.1). This is because there are too few busy periods in order for the central limit theorem to provide a good approximation. Our approximation of the standard deviation is motivated by the applicability of the central limit theorem. Because we are looking at a stochastic number of busy periods, n, the standard deviation will also be different from $\sigma = \varsigma/\sqrt{n}$ in shorter intervals. Consequently, our standard deviation approximation will have a lower accuracy.

To assess the accuracy of the standard deviation approximation in shorter intervals, we have performed additional experiments. In Table 2.2 the results are shown on the two examples for intervals ranging from 30 minutes up to 1,440 minutes. The table shows the simulated standard deviation σ, the approximated standard deviation $\hat{\sigma}$, and the relative difference between these two. The simulations have been performed 10,000 times for an accurate measurement. Two observations can be made. First, as the intervals become smaller, the standard deviation becomes larger. Second, as the intervals become smaller, the accuracy of the approximation diminishes. Both observations were explained earlier. There is also a difference between the large and the small system. The approximation of the standard deviation is more accurate on the small system. This is likely the result of a smaller busy-period length, because the offered load is less.

2.3.2 Normal Approximation

Although the relative differences of the standard deviation approximation can be quite significant for small intervals, what is more important is the accuracy of the normal approximation that uses this standard deviation approximation. As we show in this subsection, the accuracy of the resulting normal approximation is good. In total we get that the service-level distribution can be approximated as

$$\mathrm{SL} \sim \mathcal{N}(\mathbb{E}\mathrm{SL}, \hat{\sigma}^2). \tag{2.4}$$

The mean of the service-level distribution is equal to the outcome of the Erlang C formula (2.1). The standard deviation is defined by Equations (2.2) and (2.3).

	Large system			Small system		
t (minutes)	σ	$\hat{\sigma}$	Δ%	σ	$\hat{\sigma}$	Δ%
30	0.260	0.372	43.248	0.218	0.278	27.750
60	0.214	0.263	22.887	0.173	0.197	13.709
120	0.166	0.186	11.785	0.131	0.139	6.309
180	0.140	0.152	8.114	0.109	0.114	3.810
360	0.103	0.107	4.546	0.079	0.080	1.295
720	0.074	0.076	2.879	0.057	0.057	0.003
1,440	0.053	0.054	2.243	0.040	0.040	0.291

Table 2.2. Accuracy assessment of the standard deviation approximation for several interval lengths t.

There are two possible sources of error in this approximation. First, the standard deviation might not be estimated correctly. We assessed the accuracy of the standard deviation approximation in the previous subsection. Second, the normal distribution itself might not be a good distribution for the service level. We can test this.

To test the null hypothesis that a sample from the unknown service-level distribution comes from a distribution in the normal family, we perform the Lilliefors test (Lilliefors 1967). This is a goodness-of-fit test similar to the Kolmogorov-Smirnov test, with the difference that the mean and variance of the sample are used in the null hypothesis. The test statistic is

$$D = \max_x |G(x) - F(x)|,$$

where G is the empirical cumulative distribution function estimated from the sample, and F is the normal cumulative distribution function with mean and standard deviation equal to the mean and standard deviation of the sample. The null hypothesis is rejected if the test statistic is larger than the critical value.

If we perform the Lilliefors test on the two examples, we find the test statistics as shown in Table 2.3. The values D are decreasing in the interval length t. This suggests that the normal distribution becomes an appropriate distribution for the service level as the intervals become larger. However, for all intervals shown in the table, the null hypothesis is rejected at a 5% significance level.

Given that we make an error in the approximation of the standard deviation and

	Large system				Small system			
t (minutes)	D	Sim	App	Δ%	D	Sim	App	Δ%
30	0.220	0.405	0.330	18.449	0.206	0.506	0.456	9.741
60	0.180	0.503	0.470	6.533	0.147	0.578	0.561	2.981
120	0.123	0.580	0.569	1.934	0.082	0.638	0.635	0.497
180	0.089	0.617	0.613	0.730	0.069	0.667	0.667	0.024
360	0.066	0.669	0.670	0.060	0.049	0.708	0.710	0.284
720	0.047	0.709	0.710	0.122	0.036	0.738	0.740	0.266
1,440	0.032	0.738	0.738	0.096	0.025	0.760	0.761	0.159

Table 2.3. Test statistic of the normal approximation and comparison of the 0.1-quantile between the simulation and the approximation for several interval lengths t.

in the approximation by the normal distribution, we are interested in the accuracy of Equation (2.4). Motivated by the application in the next section, we assess this accuracy by comparing the 0.1-quantiles of our approximated service-level distribution with the empirical distribution based on simulations. If we denote by F^{-1} the quantile function, then we have in the former case, for $x = 0.1$,

$$F^{-1}(x) = \mathbb{E}\mathrm{SL} + \Phi^{-1}(x)\hat{\sigma},$$

where Φ^{-1} is the inverse of the standard normal cumulative distribution function. Table 2.3 lists the results of the comparison between the simulation and the approximation, together with the relative error. From these results, we can observe that the error decreases in the interval length t. This is as expected because both the standard deviation approximation and the approximation by the normal distribution become more accurate when the interval length increases. We can also see that the approximation performs very well starting from an interval length of 120–180 minutes. Therefore, when dealing with relatively slow-changing demand, the approximation is useful for such intervals. When arrival rates change significantly in a short period of time, call centers often have to divide time into smaller intervals, typically of 30 minutes. The approximation is not good in such cases.

2.4 Variability-Controlled Staffing

Staffing decisions that are made solely based on the expected value suffer from the variability in the service level. Depending on a couple of factors, it is possible

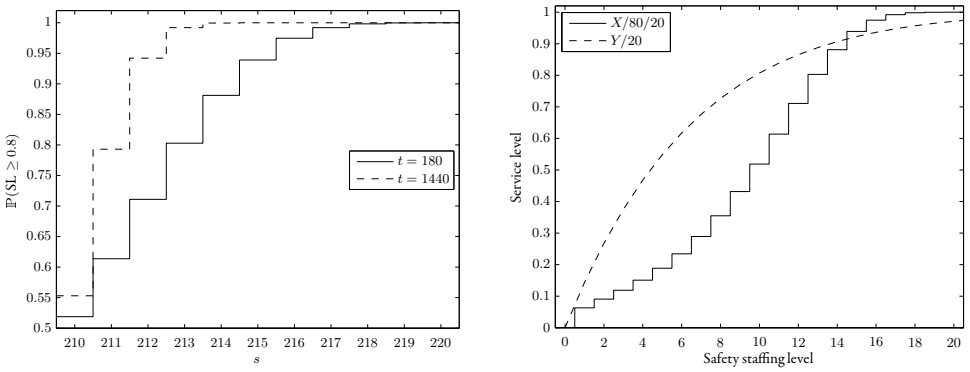


Figure 2.4. Left plot: stairs plot of the probability that the 80/20 service level will be met as a function of the number of agents for two values of t. Right plot: service level as a function of the safety staffing level. Examples based on the large system with $s^{\min} = 200$.

that the target service level will be reached only 50% of the time. These factors include, for instance, the level of aggregation and the expected service level. By making better decisions, these kinds of situations can be prevented. By taking the distribution of the service level into account, one can control the likelihood that the target service level is met.

The left plot in Figure 2.4 shows the probability that the service-level objective will be met, depending on the number of agents. From a managerial point of view, this figure is useful in two different ways. First, for a given staffing level, it could be used to show with what probability the target service level will be met. Second, for a given target, it shows the optimal staffing level. This staffing decision is based on a new service-level objective. Instead of a Y/Z service level, we now get an $X/Y/Z$ service level. This means that in $X\%$ of the intervals the target service level of Y/Z will be met. The variability-controlled staffing level $\hat{s}$ can be calculated as follows, taking 90/80/20 as an example:

$$\hat{s} = \min\left\{s \in \mathbb{N} \mid \mathbb{P}(\mathcal{N}(\mathbb{E}\mathrm{SL}, \hat{\sigma}^2) \geq 0.8) \geq 0.9\right\}. \tag{2.5}$$

Remark 2.1. The new way to do the staffing calculations in Equation (2.5) generalizes the way it is done in the Erlang C formula. When we take $t \to \infty$, we have $\hat{\sigma} \to 0$, and the approximation of the service level by the normal distribution becomes deterministic with value $\mathbb{E}\mathrm{SL}$. Then in Equation (2.5), the probability $\mathbb{P}(\mathbb{E}\mathrm{SL} \geq 0.8)$ is either one or zero. Therefore, the staffing level corresponding

to the $X/Y/Z$ service level is the same as that of the Y/Z service level for $t \to \infty$. Also, the $50/Y/Z$ service level results in the same staffing level, for all t, as the Y/Z service level. This is because the normal distribution is symmetric.

Remark 2.2. It should be noted that the staffing level corresponding to an $X/Y/Z$ target can also be achieved by a $\bar{Y}/\bar{Z}$ target, where $\bar{Y}$ or $\bar{Z}$ are possibly different from Y and Z. To illustrate this, consider, for example, the two cases in the left plot in Figure 2.4. To reach a 90/80/20 service level, we find $\hat{s} = 212$ for $t = 1{,}440$, and $\hat{s} = 215$ for $t = 180$. The same staffing levels correspond to an 84/20 and 91/20 target, respectively. The notation $X/Y/Z$ has several advantages. For instance, it is immediately clear from the target description what the likelihood is of reaching the Y/Z service level. In addition, the staffing level strongly depends on the interval length t. It is difficult to give an interpretation to $\bar{Y}/\bar{Z}$ on the likelihood X to reach Y/Z for different lengths t.

Planning according to the variability-controlled staffing level comes at higher staffing costs. The minimum number of agents needed to handle all calls in a deterministic system is $s^{\min} = \lceil \lambda/\mu \rceil$. The planning according to the traditional Y/Z service level leads to a higher number of agents $s^{Y/Z}$. The difference $s^{Y/Z} - s^{\min}$ could be interpreted as a safety staffing level to provide a higher service to the customers and is further increased to the safety staffing level $\hat{s} - s^{\min}$ according to the variability-controlled staffing of Equation (2.5). The right plot in Figure 2.4 shows the expected service level $Y/20$ and the probability X to reach the 80/20 service level as a function of the safety staffing level. To reach an 80/20 service level, a safety staffing level of 10 agents is needed. To reach this service level with a probability of 90% in an interval of $t = 180$, the safety staffing level increases to 15 agents. If the call center management includes an $X/Y/Z$ service level in their contracts, they have to consider the additional costs for these increased staffing levels in their pricing schemes.

We demonstrate the implications of our staffing approach on the staffing levels for the large and small call centers. The default staffing levels are 210 and 19 agents, respectively. Because of the observed deviation in the service level, the traditional 80/20 service level will be met only in 55.3% and 62.6% of 24-hour intervals, respectively. For different interval lengths and different target service levels, the variability-controlled staffing levels are displayed in Table 2.4. The optimal values derived via time-consuming simulations are given in parentheses. From the table, a couple of observations can be made. First, it is not surprising to see that the staffing levels increase if the traditional target service level must be met with higher

	Large system			
t (minutes)	50/80/20	90/80/20	95/80/20	99/80/20
30	210 (208)	219 (217)	220 (220)	223 (226)
60	210 (208)	217 (216)	218 (219)	220 (224)
120	210 (209)	216 (215)	217 (217)	218 (221)
180	210 (210)	215 (215)	216 (216)	217 (220)
360	210 (210)	214 (214)	214 (214)	216 (217)
720	210 (210)	213 (213)	213 (213)	214 (215)
1,440	210 (210)	212 (212)	213 (213)	213 (214)
	Small system			
t (minutes)	50/80/20	90/80/20	95/80/20	99/80/20
30	19 (18)	22 (22)	23 (23)	23 (25)
60	19 (19)	22 (21)	22 (22)	23 (24)
120	19 (19)	21 (21)	21 (22)	22 (23)
180	19 (19)	21 (21)	21 (21)	22 (22)
360	19 (19)	20 (20)	21 (21)	21 (21)
720	19 (19)	20 (20)	20 (20)	21 (21)
1,440	19 (19)	20 (20)	20 (20)	20 (20)

Table 2.4. Variability-controlled staffing levels for different target service levels and interval lengths. Optimal staffing levels are in parentheses.

probability. Second, the smaller the intervals, the more uncertainty in service level. Hence, generally more agents are needed as well. However, this does not hold for the 50/80/20 service level, because the 80/20 service level will be met with a probability higher than 50% with the default staffing levels. Third, the absolute increase in staffing levels is larger for the larger call center than it is for the smaller call center. This is because of the *law of diminishing returns* (see, e.g., Koole and Pot 2011), which states that the marginal increase in service level declines in the number of agents. An increase in expected service level is needed to ensure that the target service level is satisfied with the specified probability. Verification with simulations shows that a good amount of these staffing levels are indeed optimal. The staffing levels for the examples with $X < 99$ are optimal for $t \geq 180$, because our approximation of the service-level distribution is very accurate. In the cases

$t \leq 120$, there is a slight over- or understaffing of no more than two in our examples, except for $X = 99$. This justifies the applicability of the approximations once more.

2.5 Staffing for Nonhomogeneous Systems

The SIPP approach is a traditional approach that helps to determine performance measures and staffing levels for time-dependent systems. In these systems the parameters (essentially the arrival rate and number of agents) are dependent on the time. This is for instance denoted by the $M(t)/M/s(t)$ queueing system. From a practical point of view, the staffing levels $s(t)$ are to remain constant within a planning period, which typically has a duration of 30 minutes. In the SIPP approach, a stationary queueing model, e.g., the $M/M/s$ model, is constructed for each planning period. Each model is then independently solved for the minimum number of agents needed to meet the target service level.

In this section we show how our variability-controlled staffing approach can be integrated in the SIPP approach. To this end, we consider a real-life example of a large banking call center. Available data consist of call detail records from which we can extract, among other things, the call volumes and average service time. The call volumes are shown in the left plot in Figure 2.5, from 8.00 until 20.00. The call volumes outside this time period are negligible. The average service time turns out to be 2.5 minutes ($\mu = 0.4$).

In Tables 2.5 and 2.6 we compare the traditional approach with the variability-controlled staffing approach for different lengths of the aggregation period equal to 30 minutes, 6 hours, and 12 hours. For each approach, we report the number of staffed agents in each 30-minute interval, and we report the expected service level and the probability of meeting the service level aggregated over 30 minutes, 6 hours, and 12 hours.

When we apply the SIPP approach to this call center and model each 30-minute planning period by the $M/M/s$ queueing system, we can find the optimal staffing levels such that in each period the 80/20 target service level will be met. These staffing levels are displayed in the columns labeled 80/20 in Table 2.5. We assess the performance of this staffing approach by means of simulations. The simulations are performed using 10,000 independent replications starting from an empty system, because the call center starts empty at the beginning of the day. In the simulations we modeled the change in staffing levels from one period to the next by the so-called exhaustive discipline (see Ingolfsson 2005). This means that

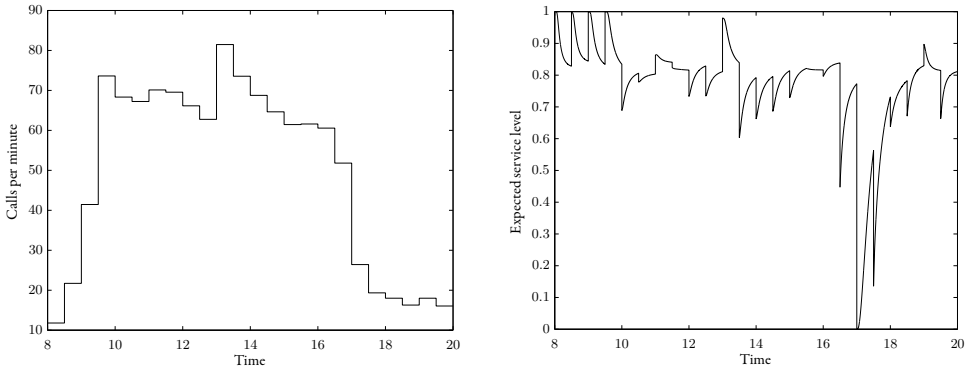


Figure 2.5. Left plot: incoming call volume by 30-minute intervals. Right plot: transient expected service level.

agents that are still serving customers will only leave as soon as they finish the call. This discipline is beneficial to the service level in periods in which the staffing level is lower than in the previous period. That the expected service level is not reached in each 30-minute period is because of the assumption of independent periods in the SIPP approach (see Stolletz 2008), where waiting customers at the end of one period are not carried over to the next period. This effect is visible in the example in Table 2.5 for periods with a significant decrease in the arrival rate compared to the former period, for example, in the period 17.00–17.30. Potentially larger queues at the end of the former period with more agents are carried over into a period with fewer agents. This leads to longer waiting times in the period with fewer agents. We can also observe this from the right plot in Figure 2.5, which shows the transient expected service level $\mathbb{E}\mathrm{SL}(t)$ for a customer arriving at time t (see Ingolfsson et al. 2007). Even though there are periods with a good average service level, the probabilities that the target service level will be met in the 30-minute periods are very low. Overall, there are 1,566.5 agent hours needed for the traditional SIPP approach without taking the variability of the service level into account.

The second part of Table 2.5 shows the results of the variability-controlled staffing according to 90/80/20 for 30-minute aggregation intervals in each 30-minute planning period, i.e., the length of the staffing period equals the length of the aggregation interval. This results in higher staffing levels and higher expected service levels. Moreover, the probabilities of reaching the desired target service

level are brought to an acceptable level. For the same reason as in the 80/20-SIPP approach, the variability-controlled SIPP approach does not reach the desired probability to reach the service level in each interval.

Usually call center managers are more interested in aggregated service levels over several hours. To integrate the length of the interval for performance measurement, we apply the variability-controlled staffing approach for the different 30-minute periods. Assume that the service levels are reported over 6-hour intervals. For each 30-minute planning period we staff according to the 90/80/20 target service level for 6-hour intervals with the arrival rate of the respective 30-minute period. This planning results in staffing decisions for short periods due to the dynamics in call volumes, but takes into account the longer intervals for performance aggregation. For aggregation intervals of 6 and 12 hours, Table 2.6 reports for each 30-minute period the staffing levels and simulation results of the expected service level and the probability that the 80/20 service level will be met. Because the staffing levels are higher than the 80/20 case and lower than the 90/80/20 30-minute case, the results are also between the two cases of Table 2.5.

Furthermore, in Tables 2.5 and 2.6 the results of the aggregated performance assessment are shown. For the aggregation of performance measures over periods with different arrival rates and staffing levels, we consider calls that start the service in the respective periods. The aggregated results show that for staffing according to 80/20 the probability to meet the 80/20 service level over the whole day is very low, with a value just above 50%. On the other hand, staffing according to 90/80/20 for 30-minute aggregation intervals gives an excessive probability. The results for staffing according to 90/80/20 for 6-hour and 12-hour aggregation intervals lie between these two extreme cases. More importantly, the probabilities to reach the service level are closer to the desired values.

The last row shows the overall agents' hours needed. The shorter the aggregation interval, the more agents are needed. In our example, the difference between the traditional approach and a 30-minute period is 61 agent hours, i.e., working with service goals for short intervals would need 3.89% more agent hours. When we compare the traditional approach with the 6-hour and 12-hour periods, we find an increase of 1.53% and 1.12% agent hours, respectively. Such analysis of additional costs is valuable in contract negotiations, where the call center management now knows the costs for a shorter aggregation interval for service-level goals.

2.6 Conclusion

In this chapter we have considered the service-level distribution beyond its expectation. When aggregated over intervals of finite length, the service level has a nonnegligible variability. Motivated by the central limit theorem, we have approximated the service-level distribution by the normal distribution. In the normal distribution the variability is characterized by the standard deviation. By means of extensive numerical experimentation, we have developed an accurate closed-form approximation for the standard deviation, depending on the length of the aggregation interval. These approximations for the service-level distribution turn out to be quite accurate, also for relatively short intervals.

Using the complete distribution of the service level, it is possible to make improved staffing decisions. Our variability-controlled staffing approach offers the possibility to control the probability that the traditional target service level is met. This results in an $X/Y/Z$ service-level objective. This means that in $X\%$ of the aggregation intervals the Y/Z target service level will be met.

Finally, we have shown, by means of an example, how our variability-controlled staffing approach could be integrated in the traditional SIPP approach to deal with time-dependent arrival rates. Because the service levels are often aggregated over several hours, we apply our approach in each small planning period to a longer aggregation interval. Although the assumptions of the SIPP approach are not justified, it is clear that our approach adds value to call center management.

A possible direction for further research could be to consider more realistic models, instead of the basic $M/M/s$ queueing system. In reality, customers are impatient and will abandon if their waiting time in the queue exceeds some (stochastic) threshold. This introduces the patience distribution as another parameter the service level depends on. Maybe abandoned customers will redial at a later time, giving rise to two more parameters: the redial probability and the redial time distribution. Furthermore, it has been shown (see, e.g., Jongbloed and Koole 2001, Avramidis et al. 2004, Brown et al. 2005) that the Poisson process cannot explain all variability in the arrival process. The arrival rate itself could therefore be modeled by a random variable. In addition, the service-time distribution differs in practice from the exponential distribution (the lognormal distribution would be more appropriate). It would be valuable if the dependence of all these characteristics on the service-level distribution could be quantified.

	80/20				90/80/20 30-minute			
Interval	s	$\mathbb{E}$SL		$\mathbb{P}(\text{SL} \geq 0.8)$	s	$\mathbb{E}$SL		$\mathbb{P}(\text{SL} \geq 0.8)$
8.00–8.30	34	0.896		0.812	37	0.971		0.970
8.30–9.00	60	0.898		0.820	64	0.975		0.975
9.00–9.30	110	0.906		0.832	115	0.974		0.965
9.30–10.00	191	0.914		0.845	198	0.983		0.980
10.00–10.30	178	0.773		0.612	184	0.946		0.916
10.30–11.00	175	0.796	0.828	0.653	181	0.954	0.959	0.931
11.00–11.30	183	0.849		0.739	189	0.968		0.955
11.30–12.00	181	0.816		0.682	187	0.959		0.941
12.00–12.30	173	0.799		0.654	178	0.945		0.912
12.30–13.00	164	0.786		0.637	170	0.951		0.924
13.00–13.30	211	0.901		0.820	218	0.980		0.972
13.30–14.00	191	0.739		0.566	197	0.929		0.884
14.00–14.30	179	0.753		0.588	185	0.945		0.913
14.30–15.00	169	0.777		0.621	175	0.953		0.927
15.00–15.30	161	0.791		0.645	166	0.946		0.917
15.30–16.00	161	0.820		0.685	167	0.962		0.943
16.00–16.30	159	0.828		0.702	164	0.957		0.934
16.30–17.00	136	0.697	0.751	0.493	142	0.918	0.928	0.869
17.00–17.30	72	0.376		0.070	76	0.657		0.291
17.30–18.00	54	0.625		0.381	57	0.875		0.786
18.00–18.30	50	0.768		0.596	54	0.953		0.935
18.30–19.00	46	0.796		0.626	49	0.941		0.914
19.00–19.30	50	0.844		0.715	54	0.965		0.959
19.30–20.00	45	0.782		0.597	48	0.932		0.897
			0.797				0.946	
				0.675				0.999
				0.382				0.984
				0.534				1.000
Agent hours	1,566.5				1,627.5			

Table 2.5. Simulation results of staffing according to 80/20 and 90/80/20 for 30-minute aggregation intervals.

	90/80/20 6-hour						90/80/20 12-hour					
Interval	s	$\mathbb{E}$SL		$\mathbb{P}(\text{SL} \geq 0.8)$			s	$\mathbb{E}$SL		$\mathbb{P}(\text{SL} \geq 0.8)$		
8.00–8.30	35	0.926	0.903	0.883	0.936	0.879	35	0.926	0.886	0.883	0.888	0.860
8.30–9.00	61	0.928		0.882			61	0.928		0.882		
9.00–9.30	112	0.940		0.903			112	0.940		0.903		
9.30–10.00	194	0.954		0.924			193	0.941		0.902		
10.00–10.30	180	0.858		0.754			180	0.849		0.735		
10.30–11.00	178	0.893		0.818			177	0.873		0.781		
11.00–11.30	185	0.919		0.859			184	0.893		0.811		
11.30–12.00	184	0.908		0.842			183	0.881		0.793		
12.00–12.30	175	0.883		0.801			174	0.854		0.747		
12.30–13.00	166	0.871		0.774			166	0.863		0.762		
13.00–13.30	214	0.951		0.919			213	0.937		0.895		
13.30–14.00	194	0.858		0.753			193	0.827		0.702		
14.00–14.30	182	0.879	0.845	0.791	0.768	0.933	181	0.846	0.860	0.731	0.671	
14.30–15.00	171	0.873		0.779			170	0.839		0.719		
15.00–15.30	163	0.877		0.787			162	0.844		0.727		
15.30–16.00	163	0.895		0.821			163	0.882		0.796		
16.00–16.30	161	0.894		0.817			160	0.878		0.788		
16.30–17.00	138	0.804		0.659			138	0.793	0.823	0.636		
17.00–17.30	73	0.470		0.111			73	0.471		0.118		
17.30–18.00	55	0.739		0.533			55	0.735		0.524		
18.00–18.30	52	0.889		0.803			51	0.848		0.727		
18.30–19.00	47	0.877		0.781			47	0.865		0.753		
19.00–19.30	52	0.925		0.881			51	0.893		0.817		
19.30–20.00	46	0.857		0.735			46	0.853		0.731		
Agent hours	1,590.5						1,584					

Table 2.6. Simulation results of staffing according to 90/80/20 for 6-hour and 12-hour aggregation intervals.

Chapter 3

Flexible Staffing with Nonstationary Arrival Rates

In this chapter we consider a multi-period staffing problem of a single-skill call center. The call center is modeled as a multi-server queue in which the staffing levels can be changed only at specific moments in time. The objective is to set the staffing levels such that a service-level constraint is met in the presence of time-varying arrival rates. We develop a Markov decision model to obtain time-dependent staffing levels for both the case where the arrival-rate function is known as well as unknown. The characteristics of the optimal policies associated with the two cases are illustrated through a numerical study based on real-life data. We show that the optimal policies provide a good balance between staffing costs and the penalty probability for not meeting the service level.

3.1 Introduction

Call centers have become the central focus of many companies, as these centers stay in direct contact with the firm's customers and form an integral part of their customer relationship management. Running a successful call center operation means managing by the numbers. One of the most important numbers in call centers is the number of agents serving incoming calls at each moment of time. Since more than two-thirds of the operating costs can be attributed to personnel, getting the right number of agents in place is critical in terms of both the offered service and the operating costs. This agent staffing problem is a complex problem in which many issues have to be taken into account, e.g., demand forecasting, variability in the call arrival patterns, quality of service, and flexibility of the workforce. We refer the reader to the comprehensive surveys in Gans et al. (2003) and Akşin et al. (2007).

In this chapter, we consider the staffing problem in a single-skill call center for a given working day. The inherent randomness in the call center, due to variability in the duration of the calls and fluctuations in the call arrival rates, makes the staff problem complex. The randomness is the root cause of deviations of the performance measures from the predicted values at the moment of planning, see, e.g., Avramidis et al. (2004), Jiménez and Koole (2004), Harrison and Zeevi (2005), Whitt (2006), Green et al. (2007), and Robbins (2007). Traditionally, most call center literature assumes known and constant mean arrival rates, mainly for the purpose of tractability. However, in addition to the usual uncertainty that is intrinsic to stochastic modeling, real call center data show that there is also uncertainty in the process parameters. Since most performance indicators are sensitive to fluctuations in the parameters (Ingolfsson et al. 2007), both types of uncertainty should be accounted for in any staffing algorithm.

A substantial body of literature has focused on the probability distribution of the arrival rates from a statistical perspective, see, e.g., Avramidis et al. (2004), Brown et al. (2004, 2005), Weinberg et al. (2007), Shen and Huang (2008a,b), Taylor (2008), Aldor-Noiman et al. (2009), and Mehrotra et al. (2010). These papers mostly deal with modeling the time-varying arrival process such that the essential features of call center arrivals are captured, e.g., a variance larger than the mean for the number of arrivals, a time-varying arrival intensity, and nonzero correlation between arrival counts in different periods.

Staffing in the presence of time-varying arrival processes was analyzed first by using the pointwise stationary approximation (PSA), see, e.g., Green and Kolesar (1991) and Green et al. (2003), in which it is assumed that the arrival rates are known, deterministic, and nonstationary. However, the PSA does not explicitly consider nonstationary behavior that may be induced by abrupt changes in the arrival rate, and it appears to perform less well in these cases. Further numerical methods have been studied by Yoo (1996), Ingolfsson et al. (2007), and Feldman et al. (2008). The first two are based on methods that solve the Chapman-Kolmogorov forward equations by using small, discrete intervals to approximate the continuously varying parameter. The latter is based on an iterative simulation-based staffing method to achieve time-stable performance.

The case of unknown nonstationary arrival rates has been studied by Jongbloed and Koole (2001), Steckley et al. (2004), Harrison and Zeevi (2005), Whitt (2006), Robbins (2007), and Liao et al. (2011). The first paper mainly focuses on the characterization of the uncertainty by providing bounds on the number of

agents needed. The second paper studies the impact of different performance measures under this uncertainty. The next two papers focus on fluid approximations to determine the number of agents. The next paper uses simulation to derive the number of agents, whereas the last paper uses a robust programming formulation. Characteristic to these papers is that they study the staffing problem under uncertainty using a fixed staffing approach. In contrast to these papers, in our problem there exists some flexibility to change the number of agents at fixed moments of the working day and thus creates more flexibility.

We model the arrival process as a nonstationary stochastic process with uncertain rates. Moreover, as is common in call centers, the call center operates under a service level that constrains the waiting time for incoming calls. The distinguishing feature of our model is that the staffing levels can only be changed at specific moments of the day, but still have to respect the service-level constraint. We assume that there is a fixed number of employees with a permanent contract and a number of flexible agents that can be changed throughout the day on specific moments. The costs of using an agent differ between the fixed and flexible agents. The objective is to find the optimal staffing level that minimizes the total call center operating costs while meeting the service-level constraint. We develop a Markov decision model that determines the optimal agent staffing policies in case the arrival-rate function is both known and unknown. We conduct a numerical study in order to illustrate the main characteristics of the optimal solutions corresponding to these approaches. In the numerical illustration, we use real call center data and show how the optimal policy balances the staffing costs and the penalty probability for not reaching the soft constraint on the service level. Furthermore, we show how the number of periods in which the staffing level can be changed affects the staffing costs.

The paper that is closest to our model is Mehrotra et al. (2010). In this paper intraday updates of the call arrival rate are also allowed. The updates are based on the cumulative number of actual arrivals and the cumulative number of expected arrivals. The ratio between these numbers is used to adjust the forecast of the next intervals. Based on the update, the new staffing levels in each period are updated using the stationary independent period-by-period (SIPP) approach (Green et al. 2001). Moreover, the performance measure used in the paper is the expected service level. In our model, we do not only look at the expected service level, but at the whole distribution of the service level, and incur a penalty when the service level at the end of the planning period is below a certain target. Hence, we need

to address the effect of a change in the number of agents on the service level at the next decision epoch. Clearly, the SIPP approach is not sufficient to address this issue. Hence, we use a dynamic programming approach to assess this impact. It is this aspect of the problem that distinguishes our model from Mehrotra et al. (2010), but also from flexible staffing models in service facilities other than call centers, e.g., Pinker and Larson (2003), Berman and Larson (2004), Bard and Purnomo (2005), Easton and Goodale (2005), and Batta et al. (2007).

The remainder of this chapter is structured as follows. In Section 3.2, we describe the call center model under consideration and formulate the associated staffing problem. In Section 3.3, we formulate our staffing algorithm for the case of both known and unknown arrival-rate functions. In Section 3.4, we conduct a numerical study to evaluate these two cases. We illustrate the impact of the number of moments at which the staffing level can be changed, and thus the benefits of flexibility in the call center. The chapter ends in Section 3.5 with concluding remarks.

3.2 Problem Formulation

Consider a call center to which customers arrive according to a nonhomogeneous Poisson process with parameter λ_t for $t \geq 0$. We assume that the call center has s_t permanent agents and f_t flexible agents at each time t, and only N workplaces available so that $s_t + f_t \leq N$ for all $t \geq 0$. If upon arrival of a new customer at time t no agent out of the $s_t + f_t$ agents is available, then the customer joins a queue with infinite buffer capacity. In the other case, the customer is directly taken into service by an idle agent and has an exponentially distributed service duration with parameter μ_t. Queued customers are served in a first-come first-served order.

The objective of the call center manager is to meet a service-level requirement by varying the number of flexible agents over the day. More precisely, divide the length of the day into m smaller intervals, each of length θ. We assume that the arrival-rate function λ_t is constant over each interval and unknown. Hence, we also take s_t to be constant over each interval. Let SL_i represent the realized service level over interval $i = 1, \ldots, m$, given by the fraction of customers that have waited no longer than the acceptable waiting time τ upon starting service within that interval. The requirement of the call center is that SL, the service level over the whole day, is at least α, where SL is computed by the average of the SL_i's weighted by the arrival counts in each interval. The decision variable of the call center

manager to achieve this requirement is the variable f_t that can only be changed at epochs determined at the start of certain intervals, namely at the start of interval $t \in \mathcal{T} = \{1, \kappa + 1, 2\kappa + 1, \ldots, m - \kappa + 1\}$, where κ is a divisor of m. Hence, the variable f_t is fixed for a longer period than λ_t and s_t, and needs to take into account the variability inherent to these variables. This is especially challenging since the arrival-rate function is not known.

The problem as described above is common in call centers. It is not realistic to assume that the number of flexible agents is changed continuously over time. The assumption that λ_t is constant over small time intervals is also not unrealistic, since this is usually the result of data estimation and forecasting procedures that approximate the true arrival-rate function when the interval length is small. We assume that the permanent agents have a cost c_1 per unit of time for each agent, and that the flexible agents cost c_2 per unit of time for each agent, with $c_2 > c_1$. Note that for any given staffing policy, one cannot guarantee that the constraint $\mathrm{SL} \geq \alpha$ is always met at the end of the day, due to randomness. When the service level at the end of the day is not met, we impose that the call center manager incurs a penalty P. We model this service-level constraint as a soft constraint. With these additional cost definitions, the problem under study becomes:

$$\min \sum_{i=1}^{m} (c_1 s_i \theta + c_2 f_i \theta) + P \mathbb{1}_{\{\mathrm{SL} < \alpha\}}$$

subject to

$$\begin{aligned} & f_t = f_{t+1} = \cdots = f_{t+\kappa-1}, && \forall t \in \mathcal{T}, \\ & s_t + f_t \leq N, && t = 1, \ldots, m, \\ & f_t \in \mathbb{N}_0, && t = 1, \ldots, m. \end{aligned}$$

3.3 Solution Approach

In order to solve the call center staffing problem, we cast the problem as a finite-horizon Markov decision problem on epochs $\mathcal{T}$. However, several simplifying approximations are required for purposes of implementation. We refer to Subsection 3.3.1 for an exact formulation that solves the problem theoretically.

Let $\mathcal{X}$ denote the state space, where at epoch $t \in \mathcal{T}$ the state $x_t \in \mathcal{X}$ denotes the service level realized up to epoch t, i.e., $x_t = \sum_{i=1}^{t-1} \tilde{\lambda}_i \mathrm{SL}_i / \sum_{i=1}^{t-1} \tilde{\lambda}_i$ for $t \in \mathcal{T}$,

where $\tilde{\lambda}_i$ is the value of λ_i derived from the observed arrival counts. Normally, the state space would be modeled by $[0, 1]$, however, we discretize the state space to $\mathcal{X} = \{0, 1/\omega, 2/\omega, \ldots, 1\}$, where the parameter ω controls how well the continuous state space is approximated. The realized service level at each epoch is rounded down to the nearest value in the new state space.

Let the action space be denoted by $\mathcal{A}_t = \{0, \ldots, N - \bar{s}_t\}$, where $\bar{s}_t = \max\{s_t, s_{t+1}, \ldots, s_{t+\kappa-1}\}$. Action $a_t \in \mathcal{A}_t$ means that the call center manager schedules $a_t = a_{t+1} = \cdots = a_{t+\kappa-1}$ flexible agents at epoch t after observing x_t.

Note that the definition of the state space is such that the Markov property does not hold. Therefore, it is impossible to give exact transition probabilities. Given that the service level x_t is known, we simulate the system to obtain the service level $x_{t+\kappa}$, given that $s_t + a_t, \ldots, s_{t+\kappa-1} + a_{t+\kappa-1}$ agents are available. We assume that at the beginning of each interval t, the system with arrival rate λ_t, service rate μ_t and $s_t + a_t$ agents has reached stationarity, which is not an unrealistic assumption when changes in the dynamics are not too severe. Starting from a steady-state situation, we apply simulations for the duration of an interval to obtain the service-level distribution. Then, by convoluting this distribution over the κ intervals, we derive the distribution for the next epoch. This approach has the advantage that we can simulate the transition probabilities up front for each combination of x_t and a_t. Hence, we can store a table with combinations of x_t and a_t that give $p_t(x_t, a_t, x_{t+\kappa})$, i.e., the probability of moving from state x_t to $x_{t+\kappa}$ when action a_t is chosen.

Finally, the direct costs are given by $c_t(x_t, a_t) = \sum_{i=t}^{t+\kappa-1}(c_1 s_i \theta + c_2 a_i \theta + P\mathbb{1}_{\{i=m\}}\mathbb{1}_{\{\text{SL}<\alpha\}})$. The first and second terms are related to the staffing of permanent and flexible agents. The last term corresponds to the penalty P that is incurred if the service level at the end of the day is not met.

The tuple $(\mathcal{X}, \mathcal{A}, p, c)$ completely describes the Markov decision process for this problem.

Note that in the problem above, the values of s_t are given. However, in practice, the values for s_t would be obtained by having an estimate of the values of λ_t for the specific day. This would typically be done in light of long-term personnel planning by using the Erlang C formula. The decision variable a_t can then be seen as short-term planning that adjusts for deviations of this estimate. It is worthwhile to mention that the use of the Erlang C formula for deriving values for s_t is not optimal in general (see Section 3.4 for some examples), but provides a good starting point for the staffing problem at hand.

In the description of our algorithm, we mention that at epoch t we define the state x_t as the realized service level up to epoch t. This service level is easy to compute, since the arrival counts in intervals $i < t$ are known. At epoch t we also need the values λ_i for $i \geq t$ to determine the optimal actions. However, these values are unknown and need to be obtained via an estimation procedure. Note that this estimation procedure can be different from the procedure used to determine the values of s_t, since the realized values up to epoch t can be used as well and provide better information on the future values of λ_t. Examples of estimation procedures can be found in, e.g., Avramidis et al. (2004), Shen and Huang (2008b), and Aldor-Noiman et al. (2009).

3.3.1 Exact Solution

In this subsection, we formulate a discrete-time Markov decision problem for our original continuous-time problem. We only discretize time into small intervals, but make no other approximation. Hence, the formulation is nearly exact for small time intervals, and thus computes nearly optimal policies for the original problem. We denote the length of a time interval by $1/\eta$, thus every $1/\eta$ time units the system is observed.

In order to model the transitions of the system after each observation, we need a large state space that contains all information to calculate the next state. Hence, define the state space $\mathcal{X}$ to consist of tuples $(n, s_c, s_d, m, z, w_1, \ldots, w_n)$. In this tuple, $n \in \mathbb{N}_0$ denotes the number of customers in the system at the time of an observation. The realized waiting times of each of the n customers at the moment of observation is given by $w_1, \ldots, w_n \in \mathbb{R}^+$. We will adopt the convention that customers in service have a waiting time of zero. Further, let $s_c \in \mathbb{N}_0$ denote the number of servers currently in use, and $s_d \in \mathbb{N}_0$ the number of servers that is desired to have. The service level can be computed by the ratio of $z \in \mathbb{N}_0$, the number of customers served within τ time units, and $m \in \mathbb{N}_0$, the number of customers served. These variables are sufficient to model the state transitions in a Markovian way. Hence, the dynamic programming backward recursion formula becomes:

$$\begin{aligned}
&\eta V_{k+1}(n, s_c, s_d, m, z, w_1, \ldots, w_n) = c_1 s_k + c_2(s_c - s_k) \\
&\quad + \lambda \mathbb{1}_{\{s_c < n\}} H_k(n+1, s_c, s_d, m, z, w_1, \ldots, w_{s_c}, w_{s_c+1} + 1/\eta, \ldots, w_n + 1/\eta, 0) \\
&\quad + \lambda \mathbb{1}_{\{s_c = n\}} H_k(n+1, s_c, s_d, m, z, w_1, \ldots, w_n, 0)
\end{aligned}$$

$$
\begin{aligned}
&+ \lambda \mathbb{1}_{\{s_c>n\}} H_k(n+1, s_c, s_d, m+1, z+1, w_1, \ldots, w_n, 0) \\
&+ \mu s_c \mathbb{1}_{\{s_c=s_d\}} \mathbb{1}_{\{s_c<n\}} \big[H_k(n-1, s_c, s_d, m+1, z + \mathbb{1}_{\{w_{s_c+1}+1/\eta<\tau\}}, \\
&\quad w_2, \ldots, w_{s_c}, 0, w_{s_c+2} + 1/\eta, \ldots, w_n + 1/\eta) \big] \\
&+ \mu s_c \mathbb{1}_{\{s_c>s_d\}} \mathbb{1}_{\{s_c<n\}} \big[H_k(n-1, s_c-1, s_d, m, z, \\
&\quad w_2, \ldots, w_{s_c}, w_{s_c+1} + 1/\eta, \ldots, w_n + 1/\eta) \big] \\
&+ \mu n \mathbb{1}_{\{s_c=s_d\}} \mathbb{1}_{\{s_c\geq n\}} \big[H_k(n-1, s_c, s_d, m, z, w_2, \ldots, w_n) \big] \\
&+ \mu n \mathbb{1}_{\{s_c>s_d\}} \mathbb{1}_{\{s_c\geq n\}} \big[H_k(n-1, s_c-1, s_d, m, z, w_2, \ldots, w_n) \big] \\
&+ \big(\eta - \lambda - \min\{n, s_c\}\mu\big) \big[\mathbb{1}_{\{s_c\geq n\}} H_k(n, s_c, s_d, m, z, w_1, \ldots, w_n) \\
&\quad + \mathbb{1}_{\{s_c<n\}} H_k(n, s_c, s_d, m, z, w_1, \ldots, w_{s_c}, w_{s_c+1} + 1/\eta, \ldots, w_n + 1/\eta) \big].
\end{aligned}
$$

The index k counts the number of intervals to go until the end of the complete period, the last interval. The first two terms describe the cost of using s_k permanent and $s_c - s_k$ flexible agents. If upon arrival, the number of servers currently in use is less than n, then there are $n - s_c$ customers in the queue. Hence, these customers add $1/\eta$ time units to their waiting time (term 3). If $s_c = n$, then everyone is in service, and the arriving customer has to wait (term 4). If $s_c > n$, then there are idle servers. Hence, an arriving customer is served immediately and satisfies the service level directly as well (term 5). The next two terms, terms 6 and 7, model the case where a customer leaves the system when there are customers waiting in the queue. The first case is where the number of servers currently in use is equal to the desired number. Hence, $1/\eta$ is added to the waiting times and s_c remains unchanged. When the customer is taken into service, then the service level is also adjusted. The second case is when s_c is higher than s_d, then additionally s_c is decreased by one and no customer is taken into service (thus, the service level is not updated either). Terms 8 and 9 model a similar situation, however, in this case there are a sufficient number of servers available so that no customer is waiting. Hence, the waiting times are not adjusted and neither is the service level. The final terms deal with the similar cases in which no event occurs within the interval. Hence, only the waiting times are updated when $s_c < n$.

In the dynamic programming backward recursion formula, we have adopted the notation H for the action operator. This action operator is equal to V for all intervals k that are not a decision epoch, i.e., $k \notin \mathcal{T}$. However, for all $k \in \mathcal{T}$, we

have that

$$H_k(n, s_c, s_d, m, z, w_1, \ldots, w_n) = \min_{l \in \mathbb{N}_0} \big\{ \{V_k(n, s_c, l, m, z, w_1, \ldots, w_n) \,|\, l < s_c\} \\ \cup \{V_k(n, l, l, m, z, w_1, \ldots, w_n) \,|\, l \geq s_c\} \big\}.$$

The first set in the minimization models the case in which the number of servers is decreased, hence s_d is adjusted. The second set models the case in which the number of servers is increased. Since this happens immediately, both s_c and s_d are set to the desired level.

Finally, we finish the model by describing what happens at the last interval. In this case, we can evaluate the realized service level and compare it to α. If the service level is not met, then a penalty of P is incurred, and otherwise no additional cost is incurred. This is given by the following equation:

$$\eta V_0(n, s_c, s_d, m, z, w_1, \ldots, w_n) = c_1 s_k + c_2(s_c - s_k) + P \mathbb{1}_{\{z/m < \alpha\}}.$$

3.4 Numerical Experiments

In this section we show the characteristics of the optimal policies by means of numerical experiments. The parameters of the experiments are based on real-life data, or otherwise chosen to represent parameters that can be found in practice. We start with an example that demonstrates the benefits of the flexibility in staffing. This example assumes a known and constant arrival rate.

Remark 3.1. In order to evaluate the optimal policies, we apply independent simulations. We mainly focus on two performance measures: the total staffing costs and the probability that a penalty is incurred if at the end of the day the service level is lower than the target. Because the penalty probability is extremely small, many simulations are necessary to obtain an accurate estimate, see, e.g., the topic of rare-event simulation in Asmussen and Glynn (2007). For instance, for a probability $p = 0.01$ and a 95% confidence interval with half-width equal to $0.1p$, the number of simulations should be at least $n = 40{,}000$. We perform $n = 1{,}000{,}000$ simulations, which can be calculated within a few seconds. We present the half-width of the 95% confidence intervals between parentheses, but omit confidence intervals for values that have a negligible half-width.

3.4.1 Constant Arrival Rate

Consider a call center with no flexibility, i.e., only a fixed number of agents are scheduled for the whole time horizon. The arrival rate is $\lambda = 3$ per minute and the service rate is $\mu = 0.2$ per minute. The acceptable waiting time is $\tau = 1/3$ minute (20 seconds). Based on these parameters, the Erlang C formula tells us that we need $s = 19$ agents in order to meet the 80% service-level target. Each agent costs $c_1 = 1$ unit per minute. Suppose that the call center operates for a time horizon of $T = 720$ minutes (12 hours). The call center starts empty, and waiting customers at the end of the time horizon are ignored.

With these parameters the costs for staffing are $C = 1 \cdot 19 \cdot 720 = 13{,}680$. However, despite the fact that the expected service level is above 80% as predicted by the Erlang C formula, simulations show that the realized service level is in many cases below 80%. With probability 0.34 the service level falls below the required target and hence a penalty is incurred. This phenomenon that the service level is not always reached in a finite-time horizon is discussed in Chapter 2. One way to deal with this problem, without flexibility, is to try a higher staffing level. With $s = 20$ the costs for staffing are increased to $C = 14{,}400$. Furthermore, the probability that a penalty is incurred is reduced to only 0.03.

We now allow flexible agents. We have a base staffing level of $s = 19$ for the whole time horizon. Each 30 minutes there is the opportunity to add flexible agents (i.e., $m = 24$, $\theta = 30$, and $\kappa = 1$), against a cost of $c_2 = 1.2$ units per minute per agent. When choosing to do so, the extra agents are immediately available. We take $N = 30$, which is sufficiently large for this example. Furthermore, we incur a sufficiently high penalty of $P = 10^6$ for failing to meet the target service level at the end of the day. The state space is discretized according to $\omega = 400$. The transition probabilities $p_t(x_t, a_t, x_{t+\kappa})$ are determined from 10,000 sample path simulations for each action a_t. By application of our method we find the optimal policy. Evaluation by means of simulations shows that the costs for staffing are $C = 14{,}192$ and that the probability of a penalty is 0.0018 $(8.4 \cdot 10^{-5})$. This is a considerable improvement in both staffing costs and penalty probability compared to staffing 20 agents with no flexibility.

The optimal policy is displayed in the left plot in Figure 3.1. This figure should be interpreted as follows. For a decision moment at time t, and given a realized service level up to t, the number of flexible agents staffed is given by the figure. The optimal policy suggests that in some cases 30 agents should be scheduled (a base

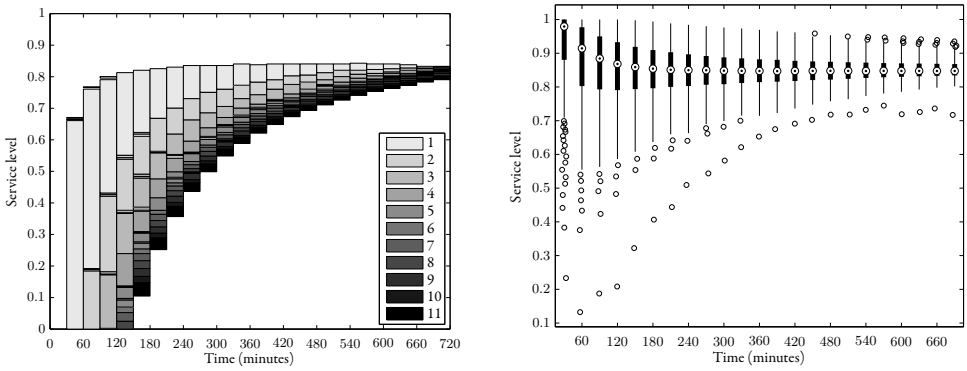


Figure 3.1. Left plot: the optimal policy for the example with a constant arrival rate. Right plot: boxplots of the service level at a decision epoch.

staffing level of 19 plus 11 flexible agents). With so many agents, the probability to reach a service level of one in a 30-minute interval is already close to one. The large white area below the curve denotes pairs of time epochs and service levels that always result in an expected service level lower than the target. Hence, no flexible agents are staffed.

The right plot in Figure 3.1 shows boxplots of the service level at a decision epoch. The small circles are outliers, where an outlier is defined as a data value more extreme than 1.5 times the interquartile range. This figure shows that the average service level nicely converges to approximately 0.85. Moreover, the variability greatly diminishes over time, and outliers are becoming more sparse.

Value of Flexibility

It is clear that staffing only a few flexible agents at the right moments keeps both the staffing costs and the penalty probability low. In the previous example we could vary the number of staffed flexible agents each 30 minutes. Allowing flexibility on this time scale might not be possible for all call centers. Also, the optimal policy was not restricted by the number of available workplaces. Therefore, we are interested in the effect of different levels of flexibility on the performance measures.

Figure 3.2 shows how the staffing costs depend on the frequency with which decisions can be made and on the number of available workplaces. The staffing costs in the left plot increase up to a maximum of $C = 15{,}840$, which is attained at staffing 22 agents for the whole day of 720 minutes. This plot shows that large

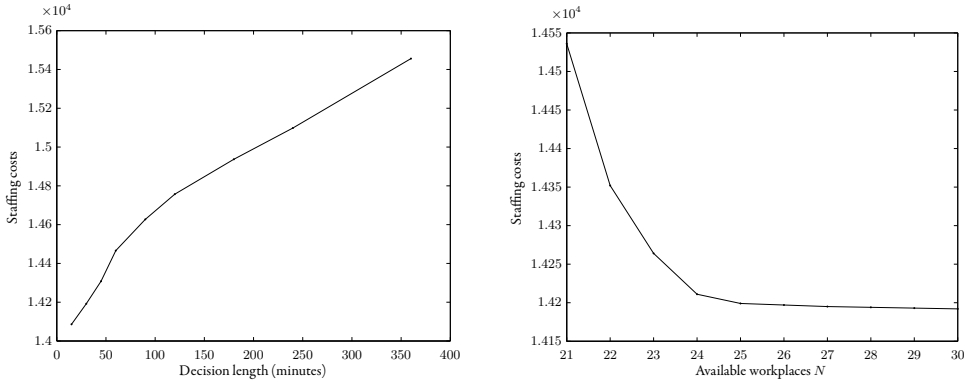


Figure 3.2. Value of flexibility.

improvements can be obtained if the call center can react on a small time scale. But there is also a significant gain if the call center can only adjust the staffing level once a day, after 360 minutes, since this reduces the staffing costs to $C = 15{,}456$. The right plot shows that most of the improvement comes from the first few flexible agents. The plot starts from a minimum of 21 workplaces, because with only 20 available workplaces there will always be a considerable penalty probability of 3%. With 21 workplaces, the staffing costs will be relatively high compared to a larger number of available workplaces, since the flexible agents are almost always used.

3.4.2 Time-Dependent Arrival Rate

We now consider a call center with a time-dependent arrival rate. We still assume that the arrival rate is known. In Figure 3.3 the typical pattern of arrivals over the day is depicted. Here we model the arrival rate as a piecewise constant function, where each interval equals 15 minutes. All other parameters related to the model remain the same. Based on the stationary Erlang C formula, we find the base staffing level in each interval such that the target will be met. These staffing levels have the same shape as the arrival rate. Performance assessment concludes that with no flexibility the staffing costs are $C = 12{,}855$ and that the probability of failing to meet the target service level at the end of the day is 0.18. When staffing one agent more in each interval, the penalty probability is reduced to 0.01, but the staffing costs are then $C = 13{,}575$.

With the opportunity to add flexible agents we can improve this situation. We assume that decisions about flexible agents can only be made each consecutive 30

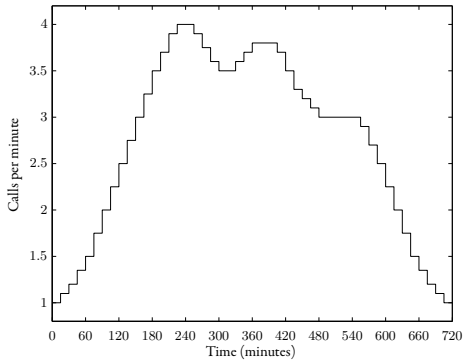


Figure 3.3. The time-dependent arrival rate.

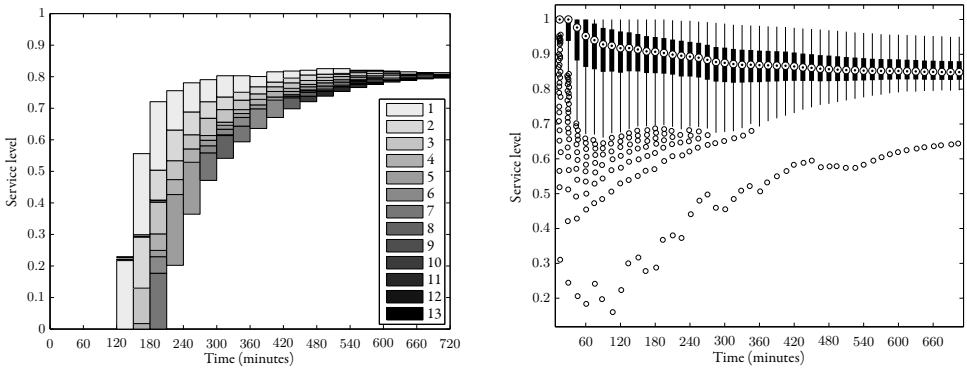


Figure 3.4. Left plot: the optimal policy for the example with a time-dependent arrival rate. Right plot: boxplots of the service level at a staffing period.

minutes, and that we have a limited number of available workplaces of $N = 30$. This implies, e.g., that we can only choose up to 5 flexible agents in the time periods $[210, 240)$ and $[240, 270)$, since there are already 25 permanent agents scheduled at $[225, 255)$. All other parameters related to our method and performance assessment remain the same. When we apply our method we find the optimal policy as shown in Figure 3.4. The corresponding costs for staffing are $C = 13{,}101$ and the probability of a penalty is 0.0062 $(1.5 \cdot 10^{-4})$. Again, this is a considerable improvement. It is astonishing to notice that this performance can be achieved by requiring on average 6.8 flexible agents at the right moments, which corresponds to only 3.4 agent hours.

The optimal policy reveals a very interesting characteristic. Until 120 minutes,

no flexible agents are needed at all. This is, of course, due to the low arrival rates at the beginning of the day, which means that the realized service level up to 120 minutes is not that important. Consequently, this provides an excellent opportunity to better estimate the arrival rate in the remainder of the day, in case the arrival rate is unknown.

The right plot in Figure 3.4 shows boxplots of the service level at the beginning of each 15-minute staffing period. Most notably from this figure is that the whiskers extending from the bottom of the boxes are becoming shorter. There are hardly any realized service levels below the 80% service-level target at the end of the day, which demonstrates that our method works well for this example.

Optimal Permanent Agents

Although the focus of this chapter is on flexible agents, we do make a short remark about the choice of the permanent agents. The flexibility of adding agents at the decision epochs can, and should, be taken into account when making a long-term planning of the permanent agents. Our method can also be used to do this. Consider for example the following heuristic approach. Start with the vector s such that $s_i = \min\{k \in \mathbb{N} \mid k > \lambda_i/\mu_i\}$ for $i = 1, \ldots, m$. That is, the number of agents in each interval is higher than the offered load, but as small as possible. Apply our method to find the policy $\pi(s)$ and the corresponding costs $C^{\pi(s)}$. The next step consists of adding a permanent agent to exactly one interval, namely the interval that will result in the largest decline in costs. Let $s + e_i$ denote the vector with an additional agent at interval i, and let $j = \arg\min_{i=1,\ldots,m} C^{\pi(s+e_i)}$. Then, if $C^{\pi(s+e_j)} < C^{\pi(s)}$, update s to $s + e_j$. Continue this iteration until no improvement can be found anymore.

We use this heuristic approach on the previous example with the time-dependent arrival rate. All parameters remain the same, with the exception of P. We increased the penalty to $P = 10^{12}$ in order to keep the penalty probability low. We find the optimal policy similar to the one in Figure 3.4. However, due to an overall decrease in the number of permanent agents, the staffing costs turn out to be much lower. They are $C = 12{,}935$, and the penalty probability is 0.0099 $(1.9 \cdot 10^{-4})$. More flexible agents are used now to reach the target service level. On average 18.5 flexible agents are needed for specific 30-minute intervals per day.

In Table 3.1, the number of permanent agents is given for each 15-minute interval. This table compares the heuristically optimal staffing levels with the base

Base	8, 9, 9, 10, 11, 12, 14, 15, 17, 18, 19, 21, 22, 23, 24, 25, 25, 24, 23, 23, 22, 22, 23, 23, 24, 24, 24, 23, 22, 21, 20, 20, 19, 19, 19, 19, 19, 19, 18, 17, 15, 14, 12, 11, 10, 9, 9, 8
Optimal	7, 7, 8, 9, 10, 11, 13, 15, 16, 17, 19, 21, 23, 24, 25, 25, 25, 25, 24, 23, 22, 22, 23, 23, 23, 23, 24, 23, 22, 21, 20, 19, 18, 18, 18, 18, 19, 18, 16, 15, 13, 11, 9, 8, 7, 6, 6, 6

Table 3.1. The number of permanent agents.

staffing levels, where in each interval the target service level will be met according to the Erlang C formula. In most of the intervals, the optimal number of permanent agents is lower, which indicates that the availability of flexible agents is better utilized when necessary. Staffing is higher in five intervals with a high arrival rate. This ensures a higher expected service level in these intervals, and possibly compensating for other intervals of lesser importance. It is interesting to note that the last couple of intervals are really understaffed. This is due to the fact that the realized service level can only be changed very limitedly near the end of the day.

3.4.3 Unknown Arrival Rate

In most practical situations the real arrival rate λ_t is not known. What is available is a best estimate $\hat{\lambda}_t$ that is estimated or forecast from historical data. It goes without saying that if this estimate is accurate ($\hat{\lambda}_t$ is close to λ_t) our method works well, in the sense that the service-level requirement will always be achieved, because it reduces to the case of a known arrival rate. What we are interested in is the performance in case the arrival-rate estimate is inaccurate.

As more information becomes available over the course of the day, our algorithm updates the arrival-rate estimate. In practice this can be done quite accurately, since large databases with historical arrival rates are available, and sophisticated updating procedures can be used (see, e.g., Shen and Huang 2008b). However, what we will show is that even with no knowledge of previous arrival rates, and therefore using a very basic updating method, our algorithm works just as well. The updating method under consideration is the historical proportion method (Shen and Huang 2008b), which works as follows. At decision epoch $t \in \mathcal{T}$ calculate the ratio R between the realized and estimated arrival rates up to t, i.e., $R = \sum_{i=1}^{t-1} \tilde{\lambda}_i / \sum_{i=1}^{t-1} \hat{\lambda}_i$. Then, update the estimate for the remainder of the day:

$\hat{\lambda}'_i = R\hat{\lambda}_i, i = t, \ldots, m$. This new estimate, together with the realized arrival rate, is then used to give an updated optimal policy.

As a result of this updating procedure, we need to evaluate the optimal policy multiple times per day. This computation takes roughly one minute to carry out, namely we update 24 times a day for a calculation that runs for approximately a few seconds. Hence, an accurate evaluation by means of extensive simulations becomes hardly doable if $n = 1{,}000{,}000$ (see also Remark 3.1). Therefore, we have to settle for less accuracy in our simulations with $n = 1{,}000$. Also, the state space is now discretized according to $\omega = 200$. As in the examples before, we take $N = 30$, a penalty of $P = 10^6$, and allow flexible agents each consecutive 30 minutes.

In our experiments, we consider several cases with respect to the pattern of the arrival rate. In the first example the estimated arrival rate is the real arrival rate multiplied by a constant scalar, $\hat{\lambda}_t = \lambda_t \cdot \beta$. In the second and third examples we correctly estimate the arrival-rate pattern, but we make a fixed under- or overestimation, $\hat{\lambda}_t = \lambda_t + \beta$, with $\beta = -0.5$ and $\beta = 0.5$. Finally, the fourth and fifth examples are examples with a wrongly estimated pattern, $\hat{\lambda}_t = \lambda_t \cdot \beta_t$, with $\beta_t = 1 - 0.005t$ and $\beta_t = 1 + 0.005t$. That is, the estimate becomes increasingly more wrong. In all examples, the true arrival rate is the one shown in Figure 3.3.

For a fair comparison between the performance of the different examples, we use the same number of permanent agents in each interval across the examples, which is the number determined by the Erlang C formula using λ_t, $\mu = 0.2$, $\tau = 1/3$, and $\alpha = 0.8$ in each interval (i.e., the base staffing levels). A reason against using the Erlang C formula with $\hat{\lambda}_t$ is that in the overestimated situations the server costs would be high and the penalty probabilities low, even without using flexible agents. Moreover, from the previous examples we have seen that the base staffing levels do require flexible agents in order to balance the server costs and the penalty probability.

The results of the experiments are shown in Table 3.2. The results for the first example are independent of β, because the β disappears in the updated estimate after the first epoch. As the day progresses, the estimate for the remainder of the day naturally becomes more accurate. Hence, this example can be seen in light of the previous example with a known and time-dependent arrival rate, though with more uncertainty. The results are also very similar. The underestimation of the arrival rate in the second example actually becomes an overestimation, because of the updating method. Therefore, more flexible agents are used resulting in higher server costs, a higher service level and a decrease in penalty probability. The third

Example	Service level	Server costs	Penalty probability
1	0.853 (0.002)	13,100 (24)	0.024 (0.009)
2	0.874 (0.002)	13,903 (31)	0.018 (0.008)
3	0.855 (0.002)	13,140 (30)	0.007 (0.005)
4	0.851 (0.002)	13,067 (28)	0.031 (0.011)
5	0.865 (0.002)	13,245 (25)	0.004 (0.004)

Table 3.2. Results of experiments with an unknown and time-varying arrival rate.

example is exactly the opposite, in the sense that for most intervals the arrival rate will be underestimated. However, an overestimation will still happen in the last ten intervals. This example shows that our method works by adapting only when the service level is too low. That the penalty probability is low is due to the overestimation in the intervals at the end. Examples four and five show results as could be expected for under- and overestimated arrival rates. That the penalty probability is not equal to zero is due to some of the simplifying approximations such as the choice of the state space, and the basic updating method.

3.5 Conclusion

In this chapter we have shown that significant improvements can be obtained by introducing flexible agents. The improvements are expressed in the form of lower staffing costs or a lower probability of failing to meet the service-level target at the end of the day, compared to the traditional approach that does not exploit this flexibility. Numerical experiments showed that our approach works remarkably well, even in the case of an unknown and time-varying arrival rate, with a forecast that is not necessarily accurate.

We model the call center as a Markov decision process in a nontraditional manner where our state variable denotes the service level as opposed to the number of customers in the system. The transition probabilities are, due to the complexities of calculating them exactly, obtained via simulations. This allows us to look further than (nonhomogeneous) Poisson arrivals and exponential service times. As more information becomes available over the course of the day, we make use of a better estimated arrival rate to update the optimal policy. In the same way, we can also update the service-time distribution. The case of agent absenteeism (e.g., a

permanent agent is scheduled to work, but did not show up) is easily handled by decreasing the number of permanent agents s_t. The absent agent will be taken care of by a flexible agent, if that turns out to be necessary.

Our approach is highly relevant to call center practice. Uncertainty in call arrivals demands flexibility from a call center to guarantee good performance without incurring excessive staffing costs. In practice, many call centers indeed have this flexibility. Flexibility in the workforce is achieved by, e.g., managers that help answering telephone calls during busy periods, or due to people that are flexible in their working hours, and can be requested to work on an ad-hoc basis with flexible contracts, such as students and agents that work from home. Additional flexibility can be obtained at the moment a shift of an agent ends and that agent can be requested to work overtime. This is practically relevant, since we observe that the demand for flexible agents increases at the end of the day, see Figure 3.4. The algorithm in this chapter exploits this flexibility in call centers in an easily implementable fashion, and therefore has the potential to be integrated in workforce management software of call centers.

Chapter 4

Service-Level Distribution of Multi-Server Queues

In this chapter we obtain the distribution of the service level after a finite time. We approach the problem in two ways. First, we approximate the distribution by studying a time-average approach using Lévy processes. Second, we provide an exact approach based on customer averages by developing a discrete-time Markov chain. We illustrate how these methods can be used in practical settings by applying the techniques to a call routing problem in a multi-site call center setting.

4.1 Introduction

It is common that the focus in service operations management is on the perceived user quality of the provided service. This user quality is commonly expressed as a service-level agreement, e.g., the fraction of customers that wait no longer than a certain acceptable waiting time should exceed a specified target (SL). Models that are used for planning in such systems generally deal with the expected service level in steady state. However, many systems operate over periods of finite length, usually no longer than 24 hours. In such small periods, the service level has a large variability, prohibiting the use of the steady-state performance measures, see Chapter 2. Therefore, it is natural to study $\mathbb{P}(\mathrm{SL}(t) \leq x)$, i.e., the probability that the service level achieved over a finite period of length t is less than or equal to $x \in [0, 1]$.

It is important to note that the randomness in the service level is due to short time-scale variations, rather than the initial distribution at the start of the interval. For example, assume that the system is stable and that the initial position is according to the stationary distribution. Common transient performance metrics, such as the waiting time of customer n, are straightforward to obtain since they

follow the same stationary distribution for every n. For the analysis of service-level variability we cannot rely on traditional transient analysis or mixing times. Despite the prime importance in practice, the performance analysis over a period of finite length has received hardly any attention in the literature.

The literature on the variability in the service level is limited to Baron and Milner (2009) and Steckley et al. (2009). They study only the limiting behavior of the service-level distribution. In the limit, the service level is shown to converge to a normal distribution, and they provide a way to compute the standard deviation of the normal deviation. In Chapter 2 a first step is provided to determine the service-level distribution in periods of finite length. Based on a normal approximation and extensive numerical experiments the standard deviation is determined in closed form.

An alternative way to study the service-level distribution is by studying the time-average equivalent, i.e., the fraction of time that the virtual waiting-time process is at or below the acceptable waiting time. Due to PASTA this converges in the limit to the customer-average service level. This is not a standard transient performance measure in the queueing literature, but has received some attention in the context of Brownian motions and diffusions. In this context, such measures are usually referred to as the occupation time. Most studies concern occupation times for Brownian motions (see, e.g., Dassios and Wu 2011) and Brownian motions with drift (see Doney and Yor 1998, Pechtl 1999). More recently, the occupation time has been derived for one-dimensional diffusions (see Pitman and Yor 2003). It is difficult to apply these results to address our problem, since the virtual waiting-time process typically has a drift (except for heavy traffic), and reflection in zero for single-server queues. For multi-server queues the behavior in zero is even more involved.

Another related performance measure may be found in reliability theory, where interval availability is considered. In reliability theory the system is often assumed to be in "up" or "down" states and the interval availability is defined as the fraction of time during a period of length t that the system is "up", see, e.g., Hosford (1960), Barlow and Hunter (1961), Rubino and Sericola (1992), or the survey Smith et al. (1997), and references therein. Related is the transient analysis of the amount of traffic generated by bursty sources, which is of interest for packet-switched communication systems (Mandjes and Van Uitert 2000). An important special case is the two-state system, where the corresponding analysis strongly builds on results of alternating renewal processes (Takács 1957). Although the underlying

models for service-level distributions are completely different, results on alternating renewal processes also provide a key starting point for our analysis in Section 4.3.

The contribution of this chapter consists of the following. First, we derive a double Laplace-Stieltjes transform of the occupation time of the virtual waiting-time process. Second, we provide the standard deviation of the occupation time in the limit. From numerical experiments we gain the insight that the distribution is typically not a normal distribution on small time scales, and large-scale systems may suffer significantly from service-level variability, which seems counter-intuitive from the principle of economies of scale. Next, we formulate an embedded Markov chain in which the service-level distribution is addressed directly, and allows the use of Markov chain machinery for analysis. Finally, we demonstrate how practical control problems can be addressed in Markov decision processes.

The remainder of this chapter is organized as follows. In Section 4.2 we introduce the model. In Section 4.3 we analyze the distribution of the time-average service level by means of a double Laplace-Stieltjes transform. In Section 4.4 we formulate an embedded Markov chain for the customer-average service-level distribution. In Section 4.5 we illustrate how the techniques can be applied in practice. We end the chapter in Section 4.6 with concluding remarks.

4.2 Model Description

Consider the $M/M/s$ queue with arrival rate λ, service rate μ, and s servers. Let τ denote the acceptable waiting time, and define $\rho = \lambda/(s\mu)$ as the load per server. It is well-known that, for $\rho < 1$, the mean service level is given by the stationary distribution of the waiting time:

$$\mathbb{E}\mathrm{SL} = \mathbb{P}(W \leq \tau) = 1 - C(s, s\rho)e^{-s\mu(1-\rho)\tau}, \tag{4.1}$$

where $C(s, s\rho)$ is the probability of delay. This waiting-time probability may be interpreted as observing the system over a very long time (long-run average) or by considering the waiting time of an arbitrary customer when the system is in steady state. In practice, one is often interested in the system behavior over finite-time intervals. Because we consider an a.s. finite number of customers with dependent waiting times, the steady-state results do not apply. In particular, these fluctuations over short intervals imply that the fraction of customers waiting no longer than τ time units will be a nondegenerate random variable, whereas $\rho \geq 1$ may give rise to a nondegenerate random variable as well.

In addition to the customer average, as in the SL definition, one may also consider the time average, i.e., the virtual waiting-time process. For single-server queues, the workload is the natural equivalent of the virtual waiting time. We like to emphasize that the performance of the virtual waiting-time process over finite intervals is a relatively unexplored area in the transient analysis of queueing systems, despite the natural fact that the human perception of performance is based on a finite sample of past performance.

4.3 Occupation Time of the Virtual Waiting-Time Process

In this section, we are interested in the fraction of time that the virtual waiting time is at or below τ, which will be called the occupation time, in line with the diffusion literature. Note that the results may also be of particular interest for single-server queues, in which case the virtual time corresponds to the workload. We consider the virtual waiting time over a finite period $(0, t]$, yielding that the time at or below τ is a random variable. For the analysis of this random variable, let $\{V(t), t \geq 0\}$ be the virtual waiting-time process. We are interested in

$$X(t) = \int_0^t \chi(u)\mathrm{d}u,$$

where $\chi(u) = \mathbb{1}_{\{V(u)\leq\tau\}}$ and $\mathbb{1}_{\{.\}}$ is the indicator function. The random variable $X(t)/t$ is thus the proportion of time that the virtual waiting-time process is at or below τ during $(0, t]$. Note that, due to PASTA, the limiting distribution of the time average $\lim_{t\to\infty} X(t)/t$ equals the (customer-average) limiting distribution of the service level.

The virtual waiting-time process in an $M/M/s$ queue is closely related to the virtual waiting-time process in an $M/M/1$ queue. In case $V(t) > 0$ all servers are occupied and the total rate of service is equal to $s\mu$, as opposed to μ in the $M/M/1$ queue. There is a difference when $V(t) = 0$. This case corresponds to an idle period, defined as the time from the moment $s - 1$ servers are occupied until the moment all servers are occupied again. In Figure 4.1 an example is shown of $V(t)$. We note that the notation will be introduced later. The jumps correspond to arrivals where the jump sizes are the times between service completions and are thus exponential with rate $s\mu$.

For the analysis of the time at or below τ during $(0, t]$, i.e., $X(t)$, we exploit that $\{\chi(t), t \geq 0\}$ is an alternating renewal process. More precisely, assume for

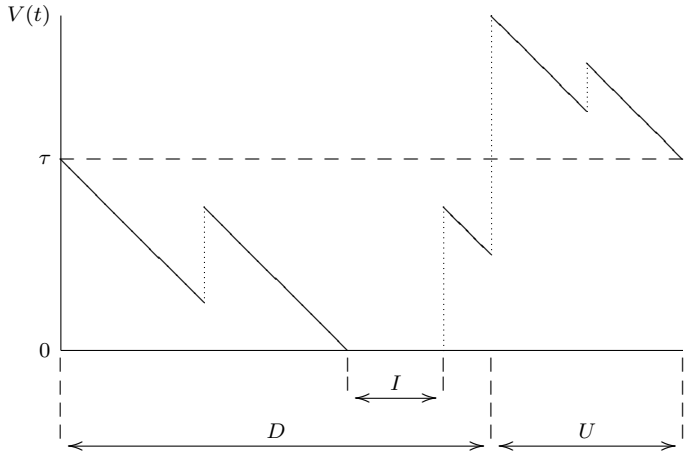


Figure 4.1. Possible realization of the virtual waiting-time process.

convenience that $V(0) = \tau + B$, with B an exponential random variable with rate $s\mu$, see Remark 4.1 for other initial conditions. Define recursively the times of upcrossings $T_{n+1}^{\text{up}} = \inf\{t > T_n^{\text{down}} : V(t) > \tau\}$ and the times of downcrossings $T_n^{\text{down}} = \inf\{t > T_n^{\text{up}} : V(t) = \tau\}$, $n = 0, 1, \ldots$ of level τ, with the convention that $T_0^{\text{up}} = 0$. The virtual waiting-time process is thus alternating between periods at or below τ and above τ according to the sequence $U_0, D_1, U_1, D_2, U_2, \ldots$, where $D_n = T_n^{\text{up}} - T_{n-1}^{\text{down}}$ and $U_n = T_n^{\text{down}} - T_n^{\text{up}}$. Since the virtual waiting-time process in an $M/M/s$ queue has continuous downcrossings and upcrossings due to exponentially distributed jumps, $\{D_n\}_{n\geq 1}$ and $\{U_n\}_{n\geq 0}$ are independent random variables. Denote their distributions by $F(\cdot)$ and $G(\cdot)$, and let $\psi_F(\cdot)$ and $\psi_G(\cdot)$ be their Laplace-Stieltjes transforms, respectively.

For the analysis of $X(t)$ we rely on the framework of Takács (1957). For the distribution of $X(t)$, we restate Theorem 1.

Theorem 4.1. *(Takács 1957, Theorem 1) The random variable $X(t)$ has the following distribution function:*

$$\Omega(x,t) = \mathbb{P}(X(t) \leq x) = \sum_{n=0}^{\infty} F_n(x) \left[G_n(t-x) - G_{n+1}(t-x)\right],$$

where F_n and G_n are the n-fold convolutions of F and G with itself.

Remark 4.1. In the framework of Takács (1957), it is assumed that the alternating renewal process starts with a regular interval of type A, whereas we measure the time that the process spends in intervals of type B. Interchanging the definitions of $F(\cdot)$ and $G(\cdot)$ provides the analysis of $1-$SL, starting in τ. Alternative initial positions may be derived by extending the analysis in Takács (1957) to allow for different first intervals. This, however, complicates the analysis and expressions, whereas the impact of the initial position is limited.

For the distribution of $X(t)$ we thus need all n-fold convolutions of F and G with itself. It is often more convenient to work with their transforms. Below, we derive the double transform of $\Omega(\cdot,\cdot)$, which is presented in Theorem 4.2. The original function can then be obtained by inversion, see Abate and Whitt (2006). Due to technical issues related to numerical inversion, we consider $\Omega(x, y+x)$, where $y = t - x$. Define the double transform

$$\phi(q,\omega) = \int_0^\infty \int_0^\infty e^{-qx} e^{-\omega y} \Omega(x, y+x) \mathrm{d}y \mathrm{d}x.$$

From Theorem 4.1 and the monotone convergence theorem we obtain an expression for the double transform:

$$\begin{aligned}
\phi(q,\omega) &= \sum_{n=0}^\infty \int_0^\infty e^{-qx} F_n(x) \mathrm{d}x \int_0^\infty e^{-\omega y} [G_n(y) - G_{n+1}(y)] \mathrm{d}y \\
&= \sum_{n=0}^\infty \frac{\psi_F(q)^n}{q} \frac{\psi_G(\omega)^n (1-\psi_G(\omega))}{\omega} \\
&= \frac{1}{q\omega} \frac{1-\psi_G(\omega)}{1-\psi_F(q)\psi_G(\omega)}.
\end{aligned}$$

We note that a similar double transform is derived in Mandjes and Van Uitert (2000).

It thus remains to specify $\psi_F(\cdot)$ and $\psi_G(\cdot)$. Let us first consider the time between upcrossing τ and the consecutive downcrossing. Due to the exponentially distributed jumps in the virtual waiting-time process, this time interval corresponds to an $M/M/s$ busy period. We define, also for later use,

$$\begin{aligned}
\eta^{\pm}(q) &= \frac{1}{2}\left(\lambda - s\mu + q \pm \delta(q)\right), \\
\delta(q) &= \sqrt{(\lambda - s\mu + q)^2 + 4s\mu q}.
\end{aligned}$$

From, e.g., Asmussen (2003), we have

$$\psi_G(q) = \frac{1}{\lambda}\left(s\mu + \eta^-(q)\right).$$

Next, we turn to the time period between a downcrossing of τ and its consecutive upcrossing, denoted by the generic random variable T_τ and having Laplace-Stieltjes transform $\psi_F(\cdot)$. For the $M/M/1$ queue, the virtual waiting-time process is a special case of a reflected Lévy process with only positive jumps. The transform of $F(\cdot)$ can then be obtained from Avram et al. (2004) and Nguyen-Ngoc and Yor (2005). For the $M/M/s$ queue, we distinguish between periods with positive virtual waiting times and idle periods. Conditioned to stay positive, the former is in fact a compound Poisson process with unit negative drift, which is a special case of a Lévy process like the $M/M/1$ queue. In the analysis below, we rely on transform results for exit times of Lévy processes.

In particular, let $\{Y_t, t \geq 0\}$ be a Lévy process with only positive jumps, let $\psi(\cdot)$ be its Lévy exponent and let $\Phi(q)$ be the largest solution of $\psi(\alpha) = q$ on the positive half axis. For the $M/M/s$ queue, the Lévy exponent reads $\psi(\alpha) = \alpha - \lambda(1 - s\mu/(s\mu + \alpha))$. Let $W^{(q)}(\cdot)$ be the q-scale function, which is defined via its transform:

$$\int_0^\infty e^{-\alpha x} W^{(q)}(x)\mathrm{d}x = \frac{1}{\psi(\alpha) - q}, \qquad \text{for } \alpha > \Phi(q),$$

and define

$$Z^{(q)}(x) = 1 + q\int_0^x W^{(q)}(y)\mathrm{d}y.$$

In the $M/M/s$ case, an explicit form for $W^{(q)}(\cdot)$ may be derived. Using the specific form of the Lévy exponent $\psi(\alpha)$ gives

$$\frac{1}{\psi(\alpha) - q} = \frac{s\mu + \alpha}{\alpha^2 + (s\mu - \lambda - q)\alpha - qs\mu} = \frac{s\mu + \alpha}{(\alpha - \eta^+(q))\,(\alpha - \eta^-(q))}.$$

Partial fraction expansion and applying Laplace inversion yields the q-scale function

$$W^{(q)}(x) = \frac{1}{\delta(q)}\left(\left(s\mu + \eta^+(q)\right)e^{\eta^+(q)x} - \left(s\mu + \eta^-(q)\right)e^{\eta^-(q)x}\right). \qquad (4.2)$$

Here, it also holds that $W^{(q)}(0) = 1$, see Kyprianou (2006, Lemma 8.6). From integrating Equation (4.2) we can directly obtain an explicit form for $Z^{(q)}(\cdot)$ as well.

In addition to T_τ, let the random variable T_0 be the first time that $V(t)$ hits zero. That is,

$$T_0 = \inf\{t > 0 : V(t) = 0\} \quad \text{and} \quad T_\tau = \inf\{t > 0 : V(t) \geq \tau\}.$$

Let $x = V(0)$ be the initial position of the reflected Lévy process $\{V(t), t \geq 0\}$. Paralleling Avram et al. (2004, Proposition 1), the transform of the time where $V(t)$ exits the interval $(0, \tau]$ below and above are, for $x \in (0, \tau]$, respectively given by

$$\begin{aligned}
\mathbb{E}_x\left[e^{-qT_0}\mathbb{1}_{\{T_0<T_\tau\}}\right] &= \frac{W^{(q)}(\tau - x)}{W^{(q)}(\tau)}, \\
\mathbb{E}_x\left[e^{-qT_\tau}\mathbb{1}_{\{T_\tau<T_0\}}\right] &= Z^{(q)}(\tau - x) - W^{(q)}(\tau - x)\frac{Z^{(q)}(\tau)}{W^{(q)}(\tau)}.
\end{aligned}$$

An application of the strong Markov property of $V(t)$ at T_0 yields

$$\begin{aligned}
\mathbb{E}_x\left[e^{-qT_\tau}\right] &= \mathbb{E}_x\left[e^{-qT_\tau}\mathbb{1}_{\{T_\tau<T_0\}}\right] + \mathbb{E}_x\left[e^{-qT_\tau}\mathbb{1}_{\{T_0<T_\tau\}}\right] \\
&= \mathbb{E}_x\left[e^{-qT_\tau}\mathbb{1}_{\{T_\tau<T_0\}}\right] + C\mathbb{E}_x\left[e^{-qT_0}\mathbb{1}_{\{T_0<T_\tau\}}\right],
\end{aligned} \tag{4.3}$$

where $C = \mathbb{E}_0\left[e^{-qT_\tau}\right]$ is the transform of the time until an upcrossing of τ starting from the moment the system becomes idle. Let $\hat{g}_{s-1}(q)$ denote the Laplace-Stieltjes transform of the idle time in the $M/M/s$ queue. Note that an idle period ends with an exponentially distributed jump in the virtual waiting-time process. Conditioning on its value, yields

$$\begin{aligned}
C &= \hat{g}_{s-1}(q)\left\{\int_0^\tau s\mu e^{-s\mu y}\mathbb{E}_y\left[e^{-qT_\tau}\right]\mathrm{d}y + e^{-s\mu\tau}\right\} \\
&= \hat{g}_{s-1}(q)\Bigg\{\int_0^\tau s\mu e^{-s\mu y}\left(\mathbb{E}_y\left[e^{-qT_\tau}\mathbb{1}_{\{T_\tau<T_0\}}\right]\right. \\
&\qquad \left. + C\mathbb{E}_y\left[e^{-qT_0}\mathbb{1}_{\{T_0<T_\tau\}}\right]\right)\mathrm{d}y + e^{-s\mu\tau}\Bigg\},
\end{aligned}$$

where the second step follows from Equation (4.3). Now, combining the results on first-exit times and the specific form of the scale function, it follows after some

lengthy but basic algebra that C is the solution of the following equation:

$$C = \hat{g}_{s-1}(q)\Bigg\{\frac{1}{\delta(q)}\left(\eta^+(q)e^{\eta^-(q)\tau} - \eta^-(q)e^{\eta^+(q)\tau}\right) + \frac{C - Z^{(q)}(\tau)}{W^{(q)}(\tau)}\frac{s\mu}{\delta(q)}\left(e^{\eta^+(q)\tau} - e^{\eta^-(q)\tau}\right)\Bigg\}.$$

The constant C can be easily obtained from the above equation. Using Equation (4.3) with $x = \tau$, the first-exit times results, and the explicit expression for the scale function again, the Laplace-Stieltjes transform is obtained, as presented in the theorem below.

Theorem 4.2. *The double transform of the distribution function* $\Omega(x, y + x)$ *is*

$$\frac{1}{q\omega}\frac{1 - \psi_G(\omega)}{1 - \psi_F(q)\psi_G(\omega)},$$

where

$$\psi_G(\omega) = \frac{1}{\lambda}\left(s\mu + \eta^-(\omega)\right)$$

$$\psi_F(q) = \tag{4.4}$$

$$1 - \frac{e^{\eta^-(q)\tau}\left[\eta^+(q)(1 - \hat{g}_{s-1}(q)) - q\right] - e^{\eta^+(q)\tau}\left[\eta^-(q)(1 - \hat{g}_{s-1}(q)) - q\right]}{e^{\eta^+(q)\tau}\left[s\mu(1 - \hat{g}_{s-1}(q)) + \eta^+(q)\right] - e^{\eta^-(q)\tau}\left[s\mu(1 - \hat{g}_{s-1}(q)) + \eta^-(q)\right]}.$$

For multi-server queues an expression for the transform of the idle time $\hat{g}_{s-1}(q)$ is given in Subsection 4.3.2. In the $M/M/1$ queue, it evidently holds that $\hat{g}_{s-1}(q) = \hat{g}_0(q) = \lambda/(\lambda + q)$. Substitution in Equation (4.4) and some rewriting yields

$$\psi_F(q) = 1 - q\frac{e^{\eta^+(q)\tau}(\mu + \eta^+(q)) - e^{\eta^-(q)\tau}(\mu + \eta^-(q))}{e^{\eta^+(q)\tau}\eta^+(q)(\mu + \eta^+(q)) - e^{\eta^-(q)\tau}\eta^-(q)(\mu + \eta^-(q))}. \tag{4.5}$$

We note that this corresponds to results in Avram et al. (2004) and Nguyen-Ngoc and Yor (2005). Specifically, set $\alpha = 0$ in the expression for $\mathbb{E}_z\left[e^{-qT_\tau - \alpha V(T_\tau)}\right]$ as given in Avram et al. (2004, Theorem 1), Nguyen-Ngoc and Yor (2005, Corollary 3), or Bekker et al. (2008, Theorem 2.1), yields

$$\mathbb{E}_z[e^{-qT_\tau}] = Z^{(q)}(\tau - z) - W^{(q)}(\tau - z)\frac{qW^{(q)}(\tau)}{W^{(q)\prime}(\tau)}.$$

Using the specific scale function (4.2) for compound Poisson processes with exponentially distributed jumps and a negative drift, the above may be rewritten as Equation (4.5).

4.3.1 Limiting Distribution

In the limit, as $t \to \infty$, the service level $\mathrm{SL}(t) = X(t)/t$ converges, after scaling, to a normal distribution. This has been observed and used as an approximation in Chapter 2 and is formally shown in Baron and Milner (2009). This limiting distribution may be directly derived from Takács (1957, Theorem 2). In the following theorem we restate the limiting result of the service-level distribution and also provide an explicit form for its variance, which has not appeared before. To do so, let $\sigma^2_{\mathrm{SL}(t)}$ be the variance of the service level and let σ^2_U and σ^2_D be the variance of times above (U) and below (D) τ, respectively. We note that the condition of Theorem 4.3 implies that $\rho < 1$. This is not required for the time-dependent analysis as presented in Theorem 4.2. For the first two moments of an idle period, represented by $\hat{g}'(0)$ and $-\hat{g}''(0)$, we refer to Subsection 4.3.2 again.

Theorem 4.3. *If $\sigma^2_U + \sigma^2_D < \infty$, then the service level is asymptotically normally distributed after scaling*

$$\lim_{t\to\infty} \mathbb{P}\left(\frac{\mathrm{SL}(t) - \mathbb{E}\mathrm{SL}}{\sigma_{\mathrm{SL}}/\sqrt{t}} \le x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} \mathrm{d}y,$$

with $\mathbb{E}\mathrm{SL}$ *given by Equation* (4.1) *and*

$$\sigma^2_{\mathrm{SL}} = (\mathbb{E}\mathrm{SL})^2(1 - \mathbb{E}\mathrm{SL}) \frac{1}{s\mu(1-\rho)} \left(\frac{\sigma^2_U}{(\mathbb{E}U)^2} + \frac{\sigma^2_D}{(\mathbb{E}D)^2}\right). \tag{4.6}$$

Here, the squared coefficients of variations of U and D are given by

$$\frac{\sigma^2_U}{(\mathbb{E}U)^2} = \frac{1+\rho}{1-\rho} \tag{4.7}$$

$$\begin{aligned}\frac{\sigma^2_D}{(\mathbb{E}D)^2} = {} & \frac{1+\rho}{1-\rho}\frac{2}{\mathbb{E}\mathrm{SL}} - \frac{1+\rho}{1-\rho} - 2\frac{1-\mathbb{E}\mathrm{SL}}{(\mathbb{E}\mathrm{SL})^2}\Bigg[s\mu\tau(1+\rho) \\ & + \frac{s\mu(1-\rho)\hat{g}'(0)}{1+s\mu(1-\rho)\hat{g}'(0)}\left(s\mu(1-\rho)\frac{\hat{g}''(0)}{2\hat{g}'(0)} + s\mu\hat{g}'(0) + \frac{1+\rho}{1-\rho}\right)\Bigg].\end{aligned} \tag{4.8}$$

Proof. Using Takács (1957, Theorem 2) it follows directly that the service level, after scaling, is asymptotically normally distributed, where $\mathbb{E}\mathrm{SL} = \mathbb{E}D/(\mathbb{E}U +$

$\mathbb{E}D$) and

$$\sigma^2_{\text{SL}} = \frac{(\mathbb{E}D)^2\sigma_U^2 + (\mathbb{E}U)^2\sigma_D^2}{(\mathbb{E}U + \mathbb{E}D)^3}.$$

The above may be rewritten as

$$\sigma^2_{\text{SL}} = \frac{(\mathbb{E}D)^2}{(\mathbb{E}U + \mathbb{E}D)^2} \times \frac{\mathbb{E}U}{\mathbb{E}U + \mathbb{E}D} \times \mathbb{E}U \left(\frac{\sigma_U^2}{(\mathbb{E}U)^2} + \frac{\sigma_D^2}{(\mathbb{E}D)^2} \right).$$

Since U is identically distributed as a busy period, we directly obtain from Asmussen (2003, p. 105) that

$$\mathbb{E}U = \frac{1}{s\mu(1-\rho)} \quad \text{and} \quad \sigma_U^2 = \frac{1+\rho}{(s\mu)^2(1-\rho)^3}.$$

Combining the above yields Equations (4.6) and (4.7).

It thus remains to specify the squared coefficient of variation of the successive times below τ. The first two moments of D can be directly derived from Equation (4.4) by differentiating with respect to q and setting $q = 0$. In particular, let $t(q)$ denote -1 times the numerator and $n(q)$ the denominator of the fraction on the right-hand side of Equation (4.4), i.e., $\psi_F(q) = 1 + t(q)/n(q)$. Noting that $t(0) = 0$, we may derive that

$$\mathbb{E}D = -\frac{t'(0)}{n(0)} \quad \text{and} \quad \frac{\sigma_D^2}{(\mathbb{E}D)^2} = \frac{n(0)t''(0)}{(t'(0))^2} - \frac{2n'(0)}{t'(0)} - 1,$$

where $\hat{g}(q) = 1 - \hat{g}_{s-1}(q)$,

$$\begin{aligned}
n(0) &= s\mu(1-\rho)e^{-s\mu(1-\rho)\tau} \\
n'(0) &= \frac{1}{1-\rho}\Big[1 + s\mu(1-\rho)\hat{g}'(0) \\
&\qquad + e^{-s\mu(1-\rho)\tau}\left(\rho - s\mu\rho(1-\rho)\tau - s\mu(1-\rho)\hat{g}'(0)\right)\Big] \\
t'(0) &= e^{-s\mu(1-\rho)\tau} - (1 + s\mu(1-\rho)\hat{g}'(0)) \\
t''(0) &= \frac{-2\tau}{1-\rho}(1 + s\mu(1-\rho)\hat{g}'(0)) - \frac{2\rho}{1-\rho}g'(0) \\
&\qquad - s\mu(1-\rho)\hat{g}''(0) - e^{-s\mu(1-\rho)\tau}\left(\frac{2\tau\rho}{1-\rho} + \frac{2\hat{g}'(0)}{1-\rho}\right).
\end{aligned}$$

Here $\hat{g}'(0)$ and $\hat{g}''(0)$ are the first and minus the second moment of an idle period. These quantities are addressed in Subsection 4.3.2 below. Using that $1 - \mathbb{E}\text{SL} = \mathbb{E}U/(\mathbb{E}U + \mathbb{E}D)$, the mean SL may be expressed in terms of the mean idle time as

$$1 - \mathbb{E}\text{SL} = \frac{1}{1 + s\mu(1-\rho)\hat{g}'(0)} e^{-s\mu(1-\rho)\tau},$$

or, equivalently, it holds that $t'(0) = \mathbb{E}\text{SL}(1 + s\mu(1-\rho)\hat{g}'(0))$. Combining the above it follows after some lengthy but straightforward calculations that the squared coefficient of variation of the successive times below τ is given by Equation (4.8). □

The variability of the service level σ^2_{SL}, as presented in Equation (4.6), has an intuitively appealing form. The term $1/(s\mu(1-\rho))$ represents the expected length of a cycle (to take the fraction of time above τ into account, we should divide by $1 - \mathbb{E}\text{SL}$), whereas the terms $\sigma^2_U/(\mathbb{E}U)^2$ and $\sigma^2_D/(\mathbb{E}D)^2$ are the squared coefficients of variation of the times above and below τ. This representation also gives insight in the service-level variability for differently sized call centers. First, we state the following lemma, which makes the economies of scale principle precise in the $M/M/s$ queueing system.

Lemma 4.1 (Economies of Scale)**.** *The expected service level increases in s, if μ and ρ are kept constant.*

Proof. The proof follows directly from Equation (4.1) and Pacheco (1994), who showed that $C(s, s\rho)$ decreases in s. □

A consequence of this lemma is that a larger system necessarily has a higher utilization to achieve the same expected service level compared to a smaller system, for equal μ and τ. Assume that μ, τ, and t are fixed and $\mathbb{E}\text{SL}$ is held constant. The expected cycle length, represented by $1/(s\mu(1-\rho))$, typically decreases due to the increased service capacity. From Equation (4.7) it is easily seen that the squared coefficient of variation of the time above τ has the opposite effect, i.e., it is strictly increasing. In the next proposition we provide a lower bound for the variance of the service level, which increases linearly with the system size. A direct consequence is that $\lim_{s\to\infty} \sigma^2_{\text{SL}}/s > 0$, revealing an important issue for large-scale systems.

Proposition 4.1. *Assume μ and τ fixed. For $\mathbb{E}\text{SL} \in (0,1)$ kept constant, we have, for some $c > 0$,*

$$\sigma^2_{\text{SL}} \geq c \times s.$$

Proof. Using $C(s, s\rho) \leq 1$ and $\mathbb{E}\text{SL} < 1$, it follows from Equation (4.1) that $s\mu(1-\rho) \leq -\log(1-\mathbb{E}\text{SL})/\tau < \infty$. As $\sigma_D^2/(\mathbb{E}D)^2 \geq 0$, we obtain from Equations (4.6) and (4.7) that

$$\frac{\sigma_{\text{SL}}^2}{s} \geq (\mathbb{E}\text{SL})^2(1-\mathbb{E}\text{SL})\frac{1}{s\mu(1-\rho)}\frac{\mu(1+\rho)}{s\mu(1-\rho)}.$$

Combining the above with $1/(s\mu(1-\rho)) \geq \tau/(-\log(1-\mathbb{E}\text{SL})) > 0$, the statement is shown. □

For single-server queues, the squared coefficient of variation of D may be considerably simplified. In particular, for $s = 1$, it holds that $\hat{g}'(0) = 1/\lambda$ and $\hat{g}''(0) = -2/\lambda^2$, and hence $1 + s\mu(1-\rho)\hat{g}'(0) = 1/\rho$. Simplifying Equation (4.8) and rewriting Equation (4.6) yields

$$\sigma_{\text{SL}}^2 = 2\mathbb{E}\text{SL}(1-\mathbb{E}\text{SL})\frac{1}{\mu(1-\rho)}\left(\frac{1+\rho}{1-\rho} - \frac{1-\mathbb{E}\text{SL}}{\mathbb{E}\text{SL}}(\mu\tau(1+\rho)+2)\right).$$

This result may be further simplified in case $\tau = 0$. Then, using $\mathbb{E}\text{SL} = 1-\rho$ and simplifying the above, yields

$$\sigma_{\text{SL}}^2 = \frac{2\rho}{\mu}. \tag{4.9}$$

4.3.2 Idle Times in Multi-Server Queues

Let the state of the $M/M/s$ queue be the number of customers in the system. Let $g_i(\cdot)$ denote the probability density function of the first-passage times from state i to state $i+1$. We are interested in an idle period $I = g_{s-1}(\cdot)$, which can be obtained recursively in the following way. The time spent in state $i < s$ is exponentially distributed with rate $\lambda + i\mu$. With probability $\lambda/(\lambda+i\mu)$ the subsequent transition is to state $i+1$, and with probability $i\mu/(\lambda+i\mu)$ the transition is to state $i-1$ $(i > 0)$. Hence, we have for $0 < i < s$,

$$g_0(t) = \lambda e^{-\lambda t},$$
$$g_i(t) = \frac{\lambda}{\lambda+i\mu}(\lambda+i\mu)e^{-(\lambda+i\mu)t} + \frac{i\mu}{\lambda+i\mu}(\lambda+i\mu)e^{-(\lambda+i\mu)t} * g_{i-1}(t) * g_i(t),$$

where $*$ means convolution. Let $\hat{g}_i(q)$ denote the Laplace-Stieltjes transform of $g_i(t)$. Then the previous expressions can be written as

$$\begin{aligned}\hat{g}_0(q) &= \frac{\lambda}{\lambda+q},\\ \hat{g}_i(q) &= \frac{\lambda}{\lambda+i\mu+q} + \frac{i\mu}{\lambda+i\mu+q}\hat{g}_{i-1}(q)\hat{g}_i(q)\\ &= \frac{\lambda}{\lambda+i\mu+q-i\mu\hat{g}_{i-1}(q)}.\end{aligned}$$

These formulas provide a recursive way to compute $\hat{g}_{s-1}(q)$ that is needed for $\psi_F(q)$ in Theorem 4.2.

Next, we determine the first and second moments of the idle times, $\mathbb{E}I = -\hat{g}'_{s-1}(0)$ and $\mathbb{E}I^2 = \hat{g}''_{s-1}(0)$ required in Theorem 4.3. Differentiating $\hat{g}_i(q)$ with respect to q gives

$$\begin{aligned}\hat{g}'_0(q) &= -\frac{\lambda}{(\lambda+q)^2},\\ \hat{g}'_i(q) &= -\frac{\lambda(1-i\mu\hat{g}'_{i-1}(q))}{(\lambda+i\mu+q-i\mu\hat{g}_{i-1}(q))^2}.\end{aligned}$$

Setting $q=0$ and noting that $\hat{g}_i(0)=1$, we may obtain $\hat{g}'_0(0) = -\frac{1}{\lambda}$ and $\hat{g}'_i(0) = -\frac{1}{\lambda}(1-i\mu\hat{g}'_{i-1}(0))$. The latter can be given in closed form, i.e.,

$$\hat{g}'_i(0) = -\sum_{k=0}^{i}\frac{i!}{(i-k)!}\frac{\mu^k}{\lambda^{k+1}},$$

which holds for $i<s$. Taking the second derivative yields

$$\begin{aligned}\hat{g}''_0(q) &= \frac{2\lambda}{(\lambda+q)^3},\\ \hat{g}''_i(q) &= \big[(\lambda+i\mu+q-i\mu\hat{g}_{i-1}(q))^2\lambda i\mu\hat{g}''_{i-1}(q)\\ &\quad +\lambda(1-i\mu\hat{g}'_{i-1}(q))2(\lambda+i\mu+q-i\mu\hat{g}_{i-1}(q))(1-i\mu\hat{g}'_{i-1}(q))\big]\\ &\quad \times(\lambda+i\mu+q-i\mu\hat{g}_{i-1}(q))^{-4}.\end{aligned}$$

Setting $q=0$ finally gives the following recursive relations

$$\begin{aligned}\hat{g}''_0(0) &= \frac{2}{\lambda^2},\\ \hat{g}''_i(0) &= \frac{1}{\lambda}i\mu\hat{g}''_{i-1}(0) + \frac{2}{\lambda^2}(1-i\mu\hat{g}'_{i-1}(0))^2.\end{aligned}$$

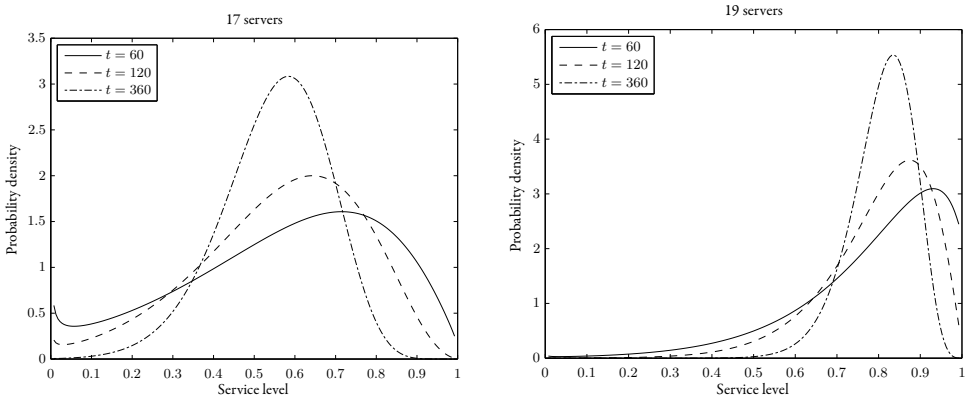


Figure 4.2. Distribution of the service level after t time units.

4.3.3 Numerical Examples

In this subsection we highlight some interesting properties of the service-level distribution. For the first example, consider the following two systems: system 1 is defined by $\lambda = 3$, $\mu = 0.2$, $s = 17$, and $\tau = 1/3$, and system 2 is defined similarly except that $s = 19$. With these parameters the expected service levels are 54.5% and 81.3% for systems 1 and 2, respectively. For various values of t we compute the distribution of the fraction of time that the virtual waiting time is at or below τ. The results are depicted in Figure 4.2. This figure shows that the variation in the service level is considerable, even for $t = 360$. Also, the distribution is far from normal when t is small, and converges to a normal distribution when t increases. It is interesting to note that the expected values of the time-average distributions are not exactly equal to the expected service levels for finite t.

As a second example we consider the limiting distribution, and illustrate how the standard deviation depends on the system parameters for a fixed expected service level. In this example we take $\mathbb{E}\mathrm{SL} = 0.8$. The left plot in Figure 4.3 shows how σ_{SL} depends on μ and τ for the $M/M/1$ queue. The right plot shows how σ_{SL} depends on s and τ for the $M/M/s$ queue with $\mu = 0.5$. The case $\tau = 0$ in the $M/M/1$ queue is interesting in the sense that σ_{SL} is decreasing in μ. This follows directly from Equation (4.9). This is different from the $M/M/s$ queue, where σ_{SL} actually starts to increase at $s = 6$ in this example. For $\tau > 0$ in the $M/M/1$ queue, σ_{SL} always seems to decrease first before it increases. In the $M/M/s$ queue, depending on τ, σ_{SL} could be decreasing first or could always be increasing. An

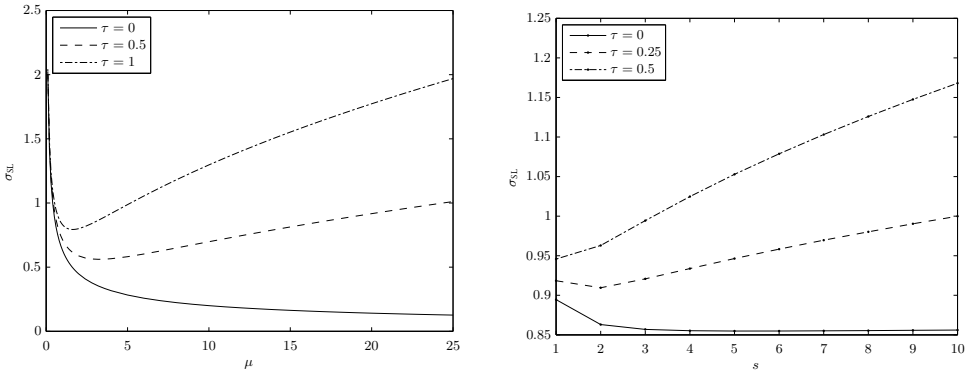


Figure 4.3. Standard deviation of the limiting service-level distribution.

interesting consequence is that by increasing the scale of the system, more variability is created, as is indicated by the lower bound of Proposition 4.1.

4.4 Embedded Markov Chain Formulation

In this section we outline an approach to determine the service-level distribution by counting the number of customers that wait no more than τ time units. Specifically, we formulate an embedded Markov chain so that it can be analyzed by techniques from the theory of Markov chains and can be utilized in an optimization framework (see also Section 4.5). Let W_n be the waiting time of the n-th customer, let B_n be the service time of the n-th customer, and let A_n be the time between the arrival moment of customer n and $n+1$. In the $G/G/1$ queue the waiting times of successive customers are related by Lindley's equation (Lindley 1952):

$$W_{n+1} = (W_n + B_n - A_n)^+.$$

This equation gives rise to the following lemma for the $M/M/1$ queue.

Lemma 4.2. *The transition probabilities are given by:*

$$\mathbb{P}(W_{n+1} = 0 \,|\, W_n = w_n) = \frac{\mu}{\lambda + \mu} e^{-\lambda w_n},$$

and, for $w_{n+1} > 0$,

$$f_{W_{n+1}\,|\,W_n=w_n}(w_{n+1}) = \begin{cases} \dfrac{\lambda\mu}{\lambda+\mu} e^{-\lambda(w_n - w_{n+1})}, & w_n \geq w_{n+1}, \\ \dfrac{\lambda\mu}{\lambda+\mu} e^{-\mu(w_{n+1} - w_n)}, & w_n < w_{n+1}. \end{cases}$$

Proof. For $W_{n+1} = 0$ we have $W_n + B_n - A_n \leq 0$, i.e., $A_n \geq W_n + B_n$. Conditioning on B_n gives

$$\mathbb{P}(W_{n+1} = 0 \,|\, W_n = w_n) = \int_0^\infty e^{-\lambda(w_n+b)} \mu e^{-\mu b} \mathrm{d}b = \frac{\mu}{\lambda+\mu} e^{-\lambda w_n}.$$

For $W_{n+1} > 0$ we have $A_n = W_n - W_{n+1} + B_n$. If $W_n \geq W_{n+1}$

$$f_{W_{n+1}\,|\,W_n=w_n}(w_{n+1}) = \int_0^\infty \lambda e^{-\lambda(w_n - w_{n+1}+b)} \mu e^{-\mu b} \mathrm{d}b = \frac{\lambda\mu}{\lambda+\mu} e^{-\lambda(w_n - w_{n+1})}.$$

If $W_n < W_{n+1}$, we require additionally that $B_n > W_{n+1} - W_n$

$$\begin{aligned} f_{W_{n+1}\,|\,W_n=w_n}(w_{n+1}) &= \int_{w_{n+1}-w_n}^\infty \lambda e^{-\lambda(w_n - w_{n+1}+b)} \mu e^{-\mu b} \mathrm{d}b \\ &= \frac{\lambda\mu}{\lambda+\mu} e^{-\mu(w_{n+1} - w_n)}. \end{aligned}$$

□

For the $M/M/s$ queue a waiting time of zero does not provide sufficient information to determine the waiting time of the next customer. The number of occupied servers is also required. Therefore, we replace a waiting time of zero by $\{-(s-1), -(s-2), \ldots, 0\}$, where $W_n = -k$ means that k servers are free after the n-th customer has joined the system, for $k = 0, 1, \ldots, s-1$. Then, using basic properties of the exponential distribution the following transition probabilities are obtained. For $i = 0, \ldots, s-1$, $j = (i-1)^+, \ldots, s-1$,

$$\mathbb{P}(W_{n+1} = -j \,|\, W_n = -i) = \prod_{k=i}^{j} \frac{(s-k)\mu}{\lambda + (s-k)\mu} \frac{\lambda}{\lambda + (s-(j+1))\mu}, \tag{4.10}$$

and for $j = 0, \ldots, s-1$, $w_n > 0$,

$$\mathbb{P}(W_{n+1} = -j \,|\, W_n = w_n) = e^{-\lambda w_n} \prod_{k=0}^{j} \frac{(s-k)\mu}{\lambda + (s-k)\mu} \frac{\lambda}{\lambda + (s-(j+1))\mu}. \tag{4.11}$$

Furthermore, from Lemma 4.2 we obtain

$$f_{W_{n+1}\,|\,W_n=w_n}(w_{n+1}) = \begin{cases} \dfrac{\lambda s\mu}{\lambda+s\mu}e^{-\lambda(w_n-w_{n+1})}, & w_n \geq w_{n+1} > 0, \\ \dfrac{\lambda s\mu}{\lambda+s\mu}e^{-s\mu(w_{n+1}-w_n)}, & 0 \leq w_n < w_{n+1}. \end{cases}$$

These transition probabilities describe the evolution of the waiting times of successive customers. In addition to the waiting time W_n, let X_n denote the number of customers that will reach the target waiting time, after the n-th arrival. The distribution of X_n is related to the waiting time in the following way: $X_n = X_{n-1} + \mathbb{1}_{\{W_n \leq \tau\}}$. Therefore, it suffices to consider the joint process $\{(X_n, W_n), n \in \mathbb{N}\}$, which is a discrete-time Markov chain. The distribution of the service level after n arrivals is then simply given by X_n/n. Denote by $p_n(x_n, w_n)$ the probability (density) that after the n-th arrival the system is in state (x_n, w_n). The following relations between the state after the n-th arrival and the $(n+1)$-th arrival can be formulated. For $j = 0, \ldots, s-1$, and for $x = 1, \ldots, n+1$,

$$\begin{aligned} p_{n+1}(x,-j) = &\sum_{i=0}^{s-1} p_n(x-1,-i)\mathbb{P}(W_{n+1} = -j \,|\, W_n = -i) \\ &+ \int_0^\infty p_n(x-1,w_n)\mathbb{P}(W_{n+1} = -j \,|\, W_n = w_n)\mathrm{d}w_n. \end{aligned}$$

For $0 < w_{n+1} \leq \tau$, and $x = 1, \ldots, n+1$,

$$\begin{aligned} p_{n+1}(x,w_{n+1}) = &\, p_n(x-1,0)f_{W_{n+1}\,|\,W_n=0}(w_{n+1}) \\ &+ \int_0^\infty p_n(x-1,w_n)f_{W_{n+1}\,|\,W_n=w_n}(w_{n+1})\mathrm{d}w_n. \end{aligned}$$

Finally, for $w_{n+1} > \tau$, and for $x = 0, \ldots, n$,

$$\begin{aligned} p_{n+1}(x,w_{n+1}) = &\, p_n(x,0)f_{W_{n+1}\,|\,W_n=0}(w_{n+1}) \\ &+ \int_0^\infty p_n(x,w_n)f_{W_{n+1}\,|\,W_n=w_n}(w_{n+1})\mathrm{d}w_n. \end{aligned}$$

The distribution of the service level after the n-th arrival, denoted by SL_n, can be computed as follows. For $x = 0, \ldots, n$,

$$\mathbb{P}(\mathrm{SL}_n = x/n) = \sum_{i=0}^{s-1} p_n(x,-i) + \int_0^\infty p_n(x,w_n)\mathrm{d}w_n.$$

4.4.1 Discretized State Space

The waiting time consists of a discrete part that counts the number of free servers when the waiting time is zero, and a positive continuous part. The continuous part of the state space can cause computational issues. To circumvent this problem we discretize the waiting time. Let $\mathcal{W} = \{-(s-1), -(s-2), \ldots, 0, \Delta, 2\Delta, \ldots, M\Delta, \widehat{M}\}$ be the discretized state space. The value $W_n = i\Delta$ has the interpretation $(i-1)\Delta < W_n \leq i\Delta$, for $i = 1, \ldots, M$, and the value $\widehat{M}$ means that $W_n > M\Delta$.

With this discretized state space the following transition probabilities can be derived. In the analysis that follows, we use that an arbitrary customer (the n-th customer) incurs the stationary waiting time. Hence, this is only possible for stable systems. See Remark 4.2 for a discussion in case $\rho \geq 1$. We divide the transition probabilities into three cases.

In the first case the n-th customer finds at least one free server, i.e., $W_n \leq 0$. Then, as in Equation (4.10), we have for $i = 0, \ldots, s-1$, $j = (i-1)^+, \ldots, s-1$,

$$\mathbb{P}(W_{n+1} = -j \,|\, W_n = -i) = \prod_{k=i}^{j} \frac{(s-k)\mu}{\lambda + (s-k)\mu} \frac{\lambda}{\lambda + (s-(j+1))\mu},$$

For $j = 1, \ldots, M$,

$$\mathbb{P}(W_{n+1} = j\Delta \,|\, W_n = 0) = \frac{\lambda}{\lambda + s\mu} \left(e^{-s\mu(j-1)\Delta} - e^{-s\mu j\Delta} \right),$$

and

$$\mathbb{P}(W_{n+1} = \widehat{M} \,|\, W_n = 0) = \frac{\lambda}{\lambda + s\mu} e^{-s\mu M\Delta}.$$

In the second case the n-th customer incurs a waiting time of $i\Delta$, $i = 1, \ldots, M$. For $j = 0, \ldots, s-1$,

$$\begin{aligned} &\mathbb{P}(W_{n+1} = -j \,|\, W_n = i\Delta) \\ &= \prod_{k=0}^{j} \frac{(s-k)\mu}{\lambda + (s-k)\mu} \frac{\lambda}{\lambda + (s-(j+1))\mu} (1-\rho) \frac{e^{-s\mu(i-1)\Delta} - e^{-s\mu i\Delta}}{e^{-s\mu(1-\rho)(i-1)\Delta} - e^{-s\mu(1-\rho)i\Delta}}. \end{aligned} \tag{4.12}$$

See below for a derivation. For $j = 1, \ldots, M$,

$$\mathbb{P}(W_{n+1} = j\Delta \,|\, W_n = i\Delta) = \begin{cases} \frac{\frac{s\mu}{\lambda+s\mu}(1-\rho)\left(e^{\lambda j\Delta} - e^{\lambda(j-1)\Delta}\right)\left(e^{-s\mu(i-1)\Delta} - e^{-s\mu i\Delta}\right)}{e^{-s\mu(1-\rho)(i-1)\Delta} - e^{-s\mu(1-\rho)i\Delta}}, & j < i, \\ \frac{2}{\lambda+s\mu}\frac{\lambda(1-e^{s\mu\Delta}) - s\mu(1-e^{\lambda\Delta})}{e^{\lambda\Delta} - e^{s\mu\Delta}}, & j = i, \\ \frac{\frac{s\mu}{\lambda+s\mu}(1-\rho)\left(e^{-s\mu(j-1)\Delta} - e^{-s\mu j\Delta}\right)\left(e^{\lambda i\Delta} - e^{\lambda(i-1)\Delta}\right)}{e^{-s\mu(1-\rho)(i-1)\Delta} - e^{-s\mu(1-\rho)i\Delta}}, & j > i, \end{cases}$$

and

$$\mathbb{P}(W_{n+1} = \widehat{M} \,|\, W_n = i\Delta) = \frac{s\mu}{\lambda+s\mu}(1-\rho)\frac{\left(e^{\lambda i\Delta} - e^{\lambda(i-1)\Delta}\right)e^{-s\mu M\Delta}}{e^{-s\mu(1-\rho)(i-1)\Delta} - e^{-s\mu(1-\rho)i\Delta}}.$$

In the third case the n-th customer incurs a waiting time of more than $M\Delta$. For $j = 0, \ldots, s-1$,

$$\mathbb{P}(W_{n+1} = -j \,|\, W_n = U) = \prod_{k=0}^{j} \frac{(s-k)\mu}{\lambda + (s-k)\mu}\frac{\lambda}{\lambda + (s-(j+1))\mu}(1-\rho)e^{-\lambda M\Delta}.$$

For $j = 1, \ldots, M$,

$$\mathbb{P}(W_{n+1} = j\Delta \,|\, W_n = \widehat{M}) = \frac{s\mu}{\lambda+s\mu}\left(e^{\lambda j\Delta} - e^{\lambda(j-1)\Delta}\right)(1-\rho)e^{-\lambda M\Delta},$$

and

$$\mathbb{P}(W_{n+1} = \widehat{M} \,|\, W_n = \widehat{M}) = \frac{2\lambda}{\lambda+s\mu}.$$

These transition probabilities follow from basic rules of conditional probabilities. As an example, we detail the analysis for the transition probabilities in Equation (4.12). All other probabilities follow in a similar way. We have

$$\mathbb{P}(W_{n+1} = -j \,|\, W_n = i\Delta) = \frac{\mathbb{P}(W_{n+1} = -j, W_n = i\Delta)}{\mathbb{P}(W_n = i\Delta)},$$

where

$$\mathbb{P}(W_n = i\Delta) = \int_{(i-1)\Delta}^{i\Delta} C(s, s\rho)s\mu(1-\rho)e^{-s\mu(1-\rho)w_n}\mathrm{d}w_n,$$

$$\mathbb{P}(W_{n+1} = -j, W_n = i\Delta)$$
$$= \int_{(i-1)\Delta}^{i\Delta} \mathbb{P}(W_{n+1} = -j \,|\, W_n = w_n)C(s, s\rho)s\mu(1-\rho)e^{-s\mu(1-\rho)w_n}\mathrm{d}w_n.$$

Using Equation (4.11) and some straightforward algebra yields Equation (4.12).

Remark 4.2. For systems with $\rho \geq 1$ such a state-space truncation is more involved. Since the system is then unstable, the probabilities $\mathbb{P}(W_{n+1} = \widehat{M} \,|\, W_n = \widehat{M})$ will become one a.s. for n large, whereas they might be smaller than one for n small. A possible solution is to let these transition probabilities be dependent on n by imposing, e.g., a linear growth in the waiting time based on a fluid limit.

We have formulated an embedded Markov chain with as state variables the discretized waiting time of the n-th customer and the number of customers that have waited no longer than τ. Note that with each additional customer in the system, the size of the state space increases. This prohibits the use of the embedded Markov chain for large values of n. However, the description of the Markov chain is explicit and enables one to use tools from the theory of Markov chains to study structural properties of the process. In particular, we refer to Rubino and Sericola (1992) and references in Smith et al. (1997).

4.5 Application to Control Problems

In this section, we illustrate how the results of Section 4.3 can be applied to controlled queueing systems. For this purpose, we study a multi-site call center problem. For many companies, the call handling function is not centralized in a single call center environment, but is dispersed across multiple sites. In such a configuration the call routing needs to be done efficiently so as to effectively use the available workforce. Aguir (2004) shows that good call routing policies result in performance close to what can be obtained through virtual resource pooling. This result is further supported by Tezcan (2008) who shows that performance optimization and load balancing results in performance approaching that of a virtual call center. However, these results were not obtained under dynamic call routing schemes, which could possibly improve performance even more significantly.

The setting for the multi-site call center in this section is as follows. We assume that calls are placed according to a Poisson process with arrival rate λ. The calls arrive at some business logic that can decide to route the call to either site 1 or site 2. The call center at site i, modeled by a multi-server queue, has s_i servers available and handles calls with a service duration that is exponentially distributed with parameter μ_i. Both call centers enforce a service level with τ minutes after a time horizon of T minutes. The objective of the business logic is to minimize the weighted sum of the probability that the service level is less than 80% at both sites. Hence, if SL_i denotes the service level at site i, then this objective is modeled by

$$w_1\mathbb{P}(\mathrm{SL}_1 < 0.8) + w_2\mathbb{P}(\mathrm{SL}_2 < 0.8),$$

where w_i is the weight representing the fraction of calls of the Poisson stream that is sent to site i. For illustration purposes, we shall fix our parameters to $\lambda = 6.5$, $\mu_1 = 0.5$, $\mu_2 = 0.25$, $s_1 = 10$, $s_2 = 15$, $\tau = 1/3$ minutes, and $T = 720$ minutes.

The routing problem stated above is rather complicated due to the form of the objective function. For example, if the objective function would be a maximization problem of the form $w_1\mathbb{E}\mathrm{SL}_1 + w_2\mathbb{E}\mathrm{SL}_2$, then the problem would be easier. In this case, one could send a fraction of calls α to site 1 and a fraction of $1 - \alpha$ to site 2. The analysis of both sites then reduces to the application of the Erlang C formula with $(\alpha\lambda, \mu_1, s_1)$ and $((1-\alpha)\lambda, \mu_2, s_2)$ as input for site 1 and site 2, respectively. Given our parameters, this would result in $\alpha^* = 0.5681$, which maximizes $\alpha\mathbb{E}\mathrm{SL}_1 + (1-\alpha)\mathbb{E}\mathrm{SL}_2$ with $\mathbb{E}\mathrm{SL}_1 = 0.8154$ and $\mathbb{E}\mathrm{SL}_2 = 0.8425$. In fact, one could even use dynamic programming to obtain a state-dependent policy. The Erlang C formula cannot be used for the model with the originally stated objective function. The Erlang C formula does not provide distributional information to calculate $\mathbb{P}(\mathrm{SL}_i < 0.8)$. However, the formulas in Section 4.3 do allow us to repeat the calculation of the previous paragraph. By using the formulas of Section 4.3, we get that $\alpha^* = 0.5622$ minimizes $\alpha\mathbb{P}(\mathrm{SL}_1 < 0.8) + (1-\alpha)\mathbb{P}(\mathrm{SL}_2 < 0.8)$ with $\mathbb{P}(\mathrm{SL}_1 < 0.8) = 0.2311$ and $\mathbb{P}(\mathrm{SL}_2 < 0.8) = 0.2784$ (with as total objective function 0.2518). Note that when $\alpha^* = 0.5681$ is used, as obtained in the case of the mean service-level objective, then the objective function has value 0.2644, which is a significant increase in total costs. Hence, our approach allows us to obtain an optimal static policy for which standard techniques cannot be applied.

In order to obtain a dynamic policy, we cast the problem into a dynamic programming framework. Note that this is not straightforward, since most dynamic

programming problems have the number of customers at each site as state variable, which does not allow us to calculate $\mathbb{P}(\text{SL}_i < 0.8)$. In order to formulate the service-level calculation as a dynamic program, we focus on one site first. Hence, we assume that arrivals to the site occur with rate λ and that individual customers are served with rate μ. Since there are s servers present in the system, the maximum rate of change that can occur is $\lambda + s\mu$.

The main idea of the approach is to record the waiting time of the customer at the head of the queue, see Koole et al. (2012), which is related to the state description in Section 4.4. The waiting times of the other customers are not stored, but can be inferred due to the fact that the arrival process is Poisson. In order to have discrete variables in the dynamic program, we discretize the waiting time into periods of expected length $1/\gamma$. The idea is to count the number of periods of exponential length having mean $1/\gamma$ that the customer at the head of the queue has been waiting before he or she enters service. In order to formulate the program, we uniformize the system with uniformization constant η (see Puterman 1994, Section 11.5). Hence, we choose η such that $\lambda + s\mu < \eta$ and that $\gamma + s\mu < \eta$. Uniformizing is equivalent to adding dummy transitions (from a state to itself) such that the rate out of each state is equal to η; then we can consider the arrival and service rates to be transition probabilities when all transitions are divided by η. Therefore, for simplicity we assume (without loss of generality) that $\eta = 1$.

Let $V_t(x, n, q)$ be a real-valued function defined on the state space $\{-s, -s + 1, \ldots\} \times \mathbb{N} \times \mathbb{N}$ at time t. The variable x denotes that there are $s + x$ servers occupied when $x \leq 0$, and it denotes the number of periods of expected length $1/\gamma$ that the customer that is at the head of the queue has been waiting. The variable n denotes the number of customers that have been taken into service, and the variable q denotes the number of customers that have entered service with a waiting time in the queue that is not above τ. The cost-to-go value function for the single multi-server queue is then defined as follows

$$V_0(x, n, q) = \mathbb{1}_{\{\frac{q}{n} < 0.8\}}.$$

This function models the criterion function when the time horizon has exceeded. The indicator function denotes $\mathbb{P}(\text{SL} < 0.8)$.

For $x < 0$, we have

$$\begin{aligned} V_t(x, n, q) = {} & \lambda V_{t-\eta}(x + 1, n + 1, q + 1) + (s + x)\mu V_{t-\eta}(x - 1, n, q) \\ & + \big(1 - \lambda - (s + x)\mu\big) V_{t-\eta}(x, n, q). \end{aligned}$$

The first term in the right-hand side represents an arrival. Since the arriving customer finds an idle server and is starting service immediately, both n and q are incremented. The next term models a departure of a customer. The last term is the dummy transition that is due to uniformization.

For $x \geq 0$, we have

$$V_t(x,n,q) = \gamma V_{t-\eta}(x+1,n,q) + s\mu \sum_{y=0}^{x} p_{x,y} V_{t-\eta}(y,n+1,q+\mathbb{1}_{\{x\leq\gamma\tau\}}) + \big(1-\gamma-s\mu\big)V_{t-\eta}(x,n,q).$$

The difference with the case $x < 0$ is that all servers are busy. Hence, the variable x now denotes the number of periods that the first customer in the queue is waiting. In contrast to the previous case, the variables n and q are only adjusted upon the start of a new service. A departure happens with rate $s\mu$. The customer that enters service at that time has been waiting for x periods accounting for a span of time equal to x/γ. Hence, the customer meets the waiting-time threshold if $x \leq \gamma\tau$. Note that the new state for x is not $x-1$. Based on the fact that the number of arrivals during a period of length x/γ has a Poisson distribution, and that the order statistics give a uniform distribution, the state y to which the process jumps is given by $p_{x,y}$ where

$$p_{x,y} = \begin{cases} 1 - \displaystyle\sum_{h=0}^{x-1} \left(\frac{\lambda}{\lambda+\gamma}\right)\left(\frac{\gamma}{\lambda+\gamma}\right)^h, & y = 0, \\ \left(\dfrac{\lambda}{\lambda+\gamma}\right)\left(\dfrac{\gamma}{\lambda+\gamma}\right)^{x-y}, & y = 1,\ldots,x. \end{cases}$$

By using the formulas of Section 4.3, we have already derived that the optimal static Bernoulli policy has $\alpha^* = 0.5622$. This policy essentially creates two independent sites for which the value function that we have derived applies to. Hence, by starting with $V_T(0,0,0)$ we can calculate the value function V^1 and V^2 for both site 1 and site 2 using the input parameters $(\alpha^*\lambda, \mu_1, s_1)$ and $((1-\alpha^*)\lambda, \mu_2, s_2)$ for each site. Note that for a dynamic policy the value function cannot be directly expressed, since $p_{x,y}$ does not hold anymore. The derivation of this probability is explicitly based on the fact that the interarrival times of calls follow a Poisson process. However, when a dynamic policy is in force, then this assumption is violated and renders the applicability of $p_{x,y}$ void. This problem can be alleviated by

also counting the number of arrivals z during a service while assuming that these z customers have arrived uniformly during the service period. In that case, $p_{x,y} = 1$ for $y = \lceil x - \frac{x}{z+1} \rceil$ (where rounding is necessary, because the waiting time has been discretized).

The approach adopted to improve the Bernoulli policy is by one-step policy improvement (see Puterman 1994, Chapter 8), which we perform in the following way. We calculate the value function $V(x, n, q, z)$ with the adjusted transition probabilities for all values of x, n, q, z for both site 1 and site 2. This value function is stored for future reference. Then the call center is simulated for performance analysis and at each decision moment, one has $x_1, n_1, q_1, z_1, x_2, n_2, q_2, z_2$ as information to base the decision on. The decision which should be based on the value function of the original problem $V_t(x_1, n_1, q_1, z_1, x_2, n_2, q_2, z_2)$, is now derived based on the approximated value function $\frac{n_1}{n_1+n_2} V_t^1(x_1, n_1, q_1, z_1) + \frac{n_2}{n_1+n_2} V_t^2(x_2, n_2, q_2, z_2)$. For example, in case there are no idle servers at both sites upon arrival at time t, one needs to determine if state $V_t(x_1, n_1, q_1, z_1 + 1, x_2, n_2, q_2, z_2)$ is better than $V_t(x_1, n_1, q_1, z_1, x_2, n_2, q_2, z_2 + 1)$. If so, routing the call to site 1 is better than routing to site 2, and vice versa. When this technique is applied to the original problem, starting with α^*, we obtain a 17% improvement over the static Bernoulli policy. Thus, the objective function has value 0.2084 with $\mathbb{P}(\mathrm{SL}_1 < 0.8) = 0.2031$ and $\mathbb{P}(\mathrm{SL}_2 < 0.8) = 0.2151$.

4.6 Conclusion

In this chapter we have studied a multi-server queue with the distinguishing feature that we obtain the distribution of the service level after a finite time. In such cases the service level is a random variable, as opposed to steady-state results in which case it is a fixed number given by the Erlang delay formula. We approach this by two methods.

First, we study the occupation time of the virtual waiting-time process, i.e., we consider the distribution of the time at or below the acceptable waiting time. This corresponds to a time average, which, due to PASTA, equals the customer average in the limit. In this method we obtain an exact expression for the double Laplace-Stieltjes transform of the time-average service-level distribution, from which the service-level distribution can be obtained using numerical inversion. From our numerical experiments it follows that the service level is highly variable and its distribution is far from normally distributed for small intervals. Such intervals

are of prime interest in call center environments, where performance is typically assessed based on time scales of minutes to hours. In the limit, the service level converges to a normal distribution. Using Laplace-Stieltjes transforms, we obtain an explicit expression for the standard deviation, revealing a potential drawback for large-scale systems showing increased variability in the service level.

Second, we develop an embedded Markov chain in which we take the number of customers that will be successfully served as state variable in addition to the waiting time of the last customer that entered the system. With this approach, we provide an explicit description of the embedded Markov chain from which the customer-average service-level distribution can be analyzed exactly using machinery for Markov chains. This Markov description may be exploited in a variety of practical optimization problems. We address a control problem for a multi-site call center, where calls are routed to one of two sites, each with their own characteristics and service-level objectives. As a first step we consider optimal static Markovian routing. Using a one-step policy improvement, both the current waiting time and the achieved service level over the past finite periods are taken into account in the state-dependent control policy. This significantly improves the performance of both sites.

Chapter 5

Performance Indicators for Call Centers with Impatience

An important feature of call center modeling is the presence of impatient customers. In this chapter, we consider single-skill call centers including customer abandonments. We study a number of different service-level definitions, including all those used in practice, and show how to explicitly compute their performance measures. Based on data from different call centers, new models are defined that extend the common Erlang A model. We show that the new models fit reality very well.

5.1 Introduction

The Erlang C model is still the most widely-used performance model in call center practice. An important property is that all delayed customers wait until they get service. In reality some calls abandon, and therefore there is a discrepancy between the Erlang C predictions and the call center reports. In the scientific community the crucial role of customer abandonments has been recognized, and a number of models has been proposed.

The most simple model including abandonments is the so-called Erlang A model. It augments the Erlang C model with patience that has an exponential distribution. Brown et al. (2005) show for a number of cases that the patience is far from exponential, and Whitt (2005) shows that the patience is the variable that is most sensitive to higher moments. It can therefore be expected that Erlang A predictions have considerable errors, and this has been confirmed in practice. One of our contributions is that we propose an extension of the Erlang A model

in which we allow for the possibility of balking. This simple extension makes the performance prediction by the queueing model much more accurate. The reason is that a relatively large proportion of the calls that get delayed abandon, and that the conditional patience distribution is approximately exponentially distributed from a certain point in time on.

Another issue is the performance measure that is used. In the Erlang C model there is an unambiguous definition of the service level: the percentage of customers that get connected before a certain acceptable waiting time (AWT). In the case of abandonments it is not immediately clear how to account for these abandonments. Different service-level definitions are used in practice, often in parallel to the abandonment percentage. In scientific work the service-level definition is often based on the virtual waiting time, i.e., the waiting time that a customer with infinite patience would experience. Note that a performance measure is only of practical use if it can also be measured in practice. For definitions based on the virtual waiting time this is not the case and therefore they are of less practical value. An important contribution is that we derive explicit expressions for several performance measures, including all measures that we have encountered in call center practice. Next, we give an overview of the literature on abandonments in call centers.

The importance of modeling abandonments in call centers is emphasized by Garnett et al. (2002), Gans et al. (2003), and Mandelbaum and Zeltyn (2009). Empirical evidence regarding abandonments in call centers can be found in Brown et al. (2005) and Feigin (2005). We refer the reader to Garnett et al. (2002), and references therein, for simple models assuming exponential patience. Garnett et al. (2002) suggest an asymptotic analysis of their Markovian abandonment model under the heavy-traffic regime. Their main result is to characterize the relation between the number of servers, the offered load, and system performance measures such as the probability of delay and the probability to abandon. This can be seen as an extension of the results of Halfin and Whitt (1981) by adding abandonments. A number of approximations for the probability to abandon are developed by Boxma and De Waal (1994). The authors have considered a multi-server queue with generally distributed service times and patience times. Brandt and Brandt (1999, 2002) consider a state-dependent Markovian multi-server queue with generally distributed patience times, in which the arrival rate depends on the number of customers in the system and in which the service rate depends on the number of busy servers. They derive the steady-state distribution of the number of customers in the system and various waiting-time distributions. The impact

of the patience distribution on the performance is studied by Mandelbaum and Zeltyn (2004). They observe an approximate linearity between the abandonment probability and the average waiting time. To analyze multi-server queues with generally distributed service times and patience times, Whitt (2005) develops an algorithm to compute approximations for standard steady-state performance measures. One of his conclusions is that the behavior of the patience distribution near the origin primarily affects the performance. Iravani and Balcıoğlu (2008a) propose two approximations that are based on scaling the single-server queue to obtain estimates for the waiting-time distributions. Other papers have treated the impatience phenomenon under various assumptions. Related studies include those by Baccelli and Hebuterne (1981), Altman and Borovkov (1997), Ward and Glynn (2003), and references therein.

Concerning the estimation of the patience distribution out of real call center data, published resources are scarce. Baccelli and Hebuterne (1981) show that an Erlang distribution with three phases works well. Kort (1983) proposes to model the patience distribution while waiting for a dial tone by the Weibull distribution. In Brown et al. (2005), it was observed that the patience distribution is not exponential as usually assumed for the call center models in the literature. However, an approximately exponential distribution was observed from a certain point in time on.

The remainder of this chapter is organized as follows. In Section 5.2, we motivate our work and give the research objectives. In Section 5.3, we conduct a statistical analysis of abandonments on real call center data. Based on this analysis, we develop in Section 5.4 a call center model and give a list of various metrics including abandonments. We then show how to explicitly compute these metrics in a convenient way. In Section 5.5, we conduct a numerical analysis in which we draw comparisons between the performance indicators. We also extend our modeling by including the important feature of retrials, and by investigating its impact on the optimal staffing level. Finally in Section 5.6, we provide some concluding remarks and directions for future research.

5.2 Context and Research Objectives

Key performance indicators (KPIs) are critical for the successful management of call centers. The right metrics identify the causes of problems and generate solutions that change the results. It is almost impossible to develop a universal set

of KPIs that will work equally well in every situation in every call center. Every business unit is different, with its unique structure and problems. Still, it is possible to formulate a set of KPIs useful for most call centers. Correct measurement of such KPIs will offer call center managers valuable information.

KPIs can be classified into two families: those that are product related and those that are process related. Product-related metrics are performance indicators mostly related to the content of the call, while process-related metrics are performance indicators that are related to call center operations.

5.2.1 Product-related metrics

The following is a list of the best-known call center product-related metrics that managers can use to improve customer experience.

First Call Resolution (FCR) FCR measures the percentage of customer issues resolved the first time. A call held waiting in the queue that ended with solving the customer issue is better than a call that got instantly connected to an agent who could not properly help the customer. A call center maintaining a good FCR rate receives a small amount of calls coming from customers who have to call back because their issue was not resolved the first time. The call center avoids therefore a significant cost due to higher call volume, increased operating expenses, and dissatisfied customers.

Turnover Turnover measures the percentage of agents who leave a call center in say a year. Turnover can be voluntary (an agent chooses to leave) or involuntary (an agent is asked to leave). A high turnover is usually considered a bad performance. It leads to high costs because of the investments in training and to problems related to agent availability.

Attendance and Punctuality Attendance is defined as an agent showing up for work on the scheduled day. Punctuality is defined as an agent showing up on time for the shift as well as being on time after breaks and lunch. One of the biggest challenges most call centers face is the control over attendance and punctuality. Low attendance and punctuality statistics can be very costly to a call center. In practice, it is common to offer incentives for good attendance and punctuality.

Contact Quality This is a common and critical customer-centric performance metric in all call centers, regardless of industry, function, and size. Top centers track contact quality as a high-level, center-wide metric, as well as an individual agent performance measure. Contact quality is typically assessed via a comprehensive evaluation form. Common quality criteria include the use of appropriate greetings and other call scripts, courtesy and professionalism, and grammar and spelling in text communication (e-mail and chat).

Customer Satisfaction Measuring customer satisfaction can be done through mail surveys and phone interviews days after the customer's interaction. Some call centers, with an advanced interactive voice response unit (IVR), survey callers immediately after the interaction occurs: customers are asked a series of questions about their interaction with the agent, their feelings about the organization, and their plans to continue doing business with the company.

5.2.2 Process-related metrics

In what follows, we give a list of the best-known classic process-related metrics used in call centers.

Probability of Blocking It measures the percentage of customers that are not able to access the center at a given time due to insufficient network facilities in place. Failure to include a blocking target allows a call center to always meet its speed-of-answer goal simply by blocking the excess calls. This damages customer accessibility and satisfaction, even though the call center appears to be doing a great job of managing the queue.

Probability of Abandonment It measures the percentage of customers that abandon the queue while waiting, i.e., leave the system without service. Time before abandonment is customer specific. However, a call center can control abandonments by controlling the waiting time in the queue (which in turn affects abandonments). Abandonment is a measure associated with interactive channels, especially calls and web chat.

Short Abandonments The short-abandonment statistic represents the total number of customers that abandon before a specified short-abandonment time.

Callers may abandon quickly for many reasons, for example, selecting an incorrect option in the IVR. Short abandonments, in contrast to regular abandonments, are not considered a sign of bad service. Therefore, many call centers count short abandonments differently, for example by not counting them at all.

Service Level, Average Speed of Answer, Longest Delay in Queue The service level measures the percentage of calls answered within a specified time limit. It is the most common process-related metric used in call centers. It is typically stated as X percent of calls handled in Y seconds or less. The average speed of answer (ASA) represents the average waiting time in the queue of all calls in a given period. Another delay measure is how long the oldest call in the queue has been waiting: the longest delay in queue (LDQ). A number of call centers use real-time LDQ to indicate when more staff need to be made immediately available. Historical data of LDQ can be used to indicate the "worst-case" experience of a customer over a period of time.

Agent Occupancy Agent occupancy is the measure of the actual time an agent is busy on customer contacts compared with the available or idle time, calculated by dividing workload hours by staff hours. Occupancy is an important measure of how well the call center has scheduled its staff and how efficiently it is using its resources.

5.2.3 Motivation

There is no single perfect or complete list of performance metrics for all call centers. On the one hand, basing an entire service strategy on the number of calls handled per hour or on the average speed of answer will inevitably lead to shortcomings in the quality. On the other hand, focusing too strongly on quality metrics while disregarding process-related measurements can still have an adverse effect on customer experience.

We focus on metrics related to queueing delays. These are classic process-related metrics that lie at the heart of effective call center and customer relations management. They are the clearest indication of what customers experience when they attempt to reach the call center. We refer the reader to Cleveland and Mayben (1997) for more details and discussions on process-related metrics. In this chapter,

we give particular attention to process-related metrics that include abandonments. The feature of customer abandonment is a crucial point in call centers.

One important point has to be clarified before impatience can be included in queueing models. That is, we need additional information concerning the patience, the willingness to wait until service commences. Similarly as for the input of the Erlang C model, the patience has to be determined from historical data. However, a number such as the average patience cannot be determined by simply averaging over the abandonment times. Indeed, the time at which other calls got connected tells us something about their patience, which should be taken into account. Statistical techniques exist to deal with these so-called censored data. Not using these methods can lead to a significant underestimation of patience, because the abandonments occur mostly among the very impatient customers. In Section 5.3, we conduct a statistical analysis on real call center data in order to characterize the statistical distribution of times before abandonments.

An advantage of Erlang C is that service-level expressions are relatively simple and easy to calculate. By taking abandonments into account, the computations become more difficult. Moreover, even when patience times are assumed to have an exponential distribution (the Erlang A model), there exist only expressions for some metrics, such as the conditional waiting time given service. In this chapter, we give a comprehensive list of the metrics including abandonments, and explicitly derive the expressions for the probability distributions of these metrics. By doing so, we obtain existing results and derive new results, such as the conditional waiting time given service of the customers who do not have short patience times.

5.3 Statistical Analysis and Modeling of Abandonments

Call center data usually consist of very detailed records about the flow of calls through the call center. It is necessary to analyze these data in order to obtain estimates for the model parameters. Moreover, it is also important to validate the modeling assumptions, a step that is often overlooked. To analyze the patience, we need to know how long customers have spent waiting, and whether an abandonment occurred at the end of the waiting time. From customers that have abandoned, we know exactly what their patience is. However, from customers that did not abandon (but received service), we only know that their patience is greater than the time they have waited. To be more precise, we observe the minimum of the patience and the virtual waiting time, and we also know which one we observe.

This is called right-censored data. Techniques exist to deal with censored data, one of which is the Kaplan-Meier estimator (see Kaplan and Meier 1958).

In our statistical analysis, we use data obtained from several real call centers. The data originate from a large banking call center located in the US, from a bank located in the Netherlands, and from a Dutch university medical center. Furthermore, we make use of the Anonymous Bank call center data available at *http://iew3.technion.ac.il/serveng/callcenterdata*.

To show the significance of uncensoring the data, consider the following example. On average 20 calls per minute arrive to a call center, and the average handling time is 5 minutes. To reach the target of 80% of the calls answered within 20 seconds, 108 agents should be scheduled according to the Erlang C formula. On a particular data set, the uncensored average patience turns out to be 780 seconds. When we apply the Erlang A model, we need 106 agents to reach the target. However, the censored average patience was 100 seconds. Using this number as the average patience, the Erlang A formula suggests an insufficient number of 95 agents. With 95 agents the real service level will only be 30%. This example also demonstrates that the Erlang C model can lead to erroneous results, since it does not predict any abandonment.

The result of the Kaplan-Meier estimator is the empirical cumulative distribution function $F(t)$ of the patience. By taking the derivative we can obtain the probability density function $f(t)$, and the hazard rate $h(t) = f(t)/(1 - F(t))$. In Figure 5.1 several empirical hazard rates are displayed. This figure also shows the hazard rates of a hyperexponential distribution, which will be discussed shortly. The empirical hazard rates are smoothed three times using a moving-average filter with a span of five, to produce better-looking lines. The patience on all four data sets can, for the most part, be characterized in the same way. In the first couple of seconds the hazard rate is high, indicating very impatient customers who are not willing to wait at all. The hazard rate quickly becomes constant thereafter, which suggests that the patience from then on is exponential. Data sets 1 and 3 show additionally several peaks in the neighborhood of sixty seconds. This is because of delay announcements in the call handling system, that actually increase the likelihood of abandoning. We do not directly model delay announcements because this property is not shared by all call centers. We refer to Jouini et al. (2011a), who study the impact of announcing delays in a setting of a single customer class with Markovian abandonments.

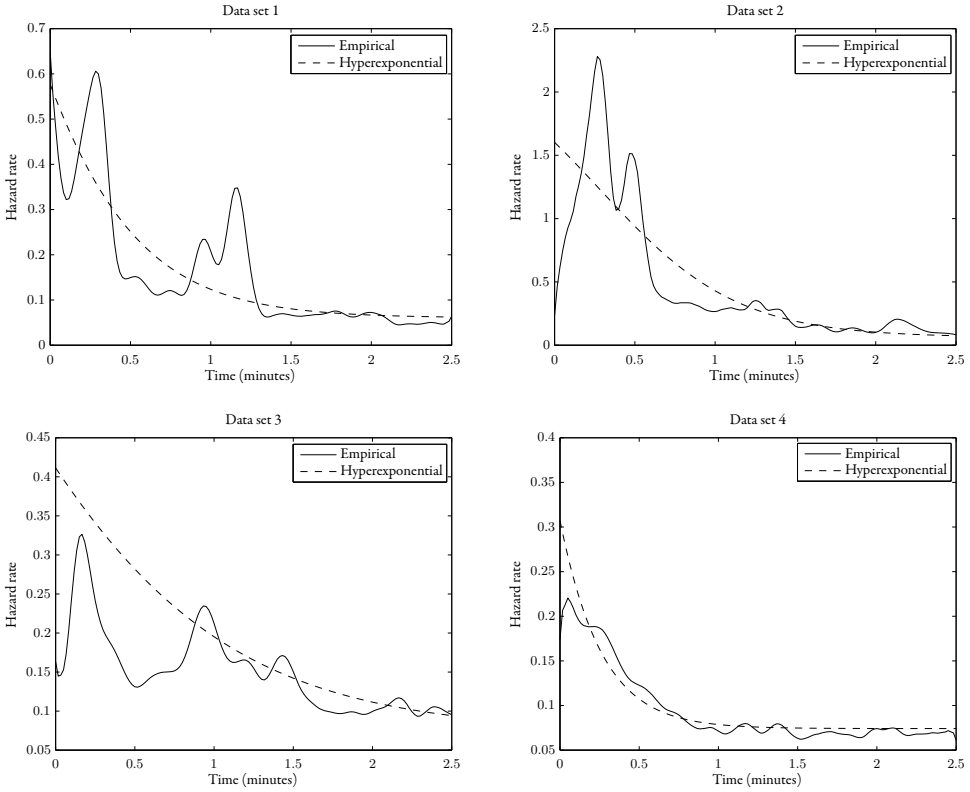


Figure 5.1. Hazard rates of the patience of four different data sets.

Model 1

A way to model this customer behavior is to extend Erlang A by including the possibility of balking. Let T denote the random variable measuring the patience times. The distribution of T consists of a discrete mass at zero corresponding to very impatient customers, and a remaining exponential distribution for customers with a positive patience. We denote by α the probability that a customer, arriving to a busy system, will immediately balk. This feature models a nonnegligible portion of the customers who immediately hang up once they know that they have to wait for service. On the other hand, with probability $1 - \alpha$, customers who find a busy system will accept to join the queue. For these customers, the patience thresholds are independent and exponentially distributed with rate γ. Hence, the cumulative

	Model 1		Model 2		
Data set	α	γ	p	γ_1	γ_2
1	0.1866	0.0656	0.2222	2.3843	0.0603
2	0.4626	0.1625	0.6593	2.3986	0.0617
3	0.1968	0.0864	0.2734	1.3100	0.0735
4	0.0506	0.0755	0.0583	4.0780	0.0742

Table 5.1. Parameters of the two models for the four data sets. The unit of time is minute.

distribution function is

$$F_T(t) = \alpha + (1-\alpha)(1-e^{-\gamma t}),$$

for $t \geq 0$.

Model 2

Another way to model customers' patience is by the hyperexponential distribution with two phases. The hyperexponential distribution is a mixture of two exponential distributions such that with probability p it is exponential with parameter γ_1 and with probability $1-p$ it is exponential with parameter γ_2. If T is hyperexponential, its cumulative distribution function F_T is given by

$$F_T(t) = p(1-e^{-\gamma_1 t}) + (1-p)(1-e^{-\gamma_2 t}),$$

for $t \geq 0$. By inspecting Figure 5.1, it seems that the hyperexponential distribution fits the empirical patience very well. The parameters of the random variable T are obtained by minimizing the mean squared error between $F(t)$ and $F_T(t)$. Table 5.1 lists these parameters for both models. From the figure we can deduce the following. Data set 4 is the perfect example of hyperexponential patience. The empirical hazard rate is approximately nonincreasing, and the hazard rate of the hyperexponential distribution follows it very closely. The fits on data sets 1 and 2 also look reasonable, even though the empirical hazard rate starts out low for the first 0.25 minutes on data set 2. This could be explained by a welcome message that is played at the start of joining the queue. The empirical hazard rate is overestimated in the first minute on data set 3, but this fit is close afterwards.

	Hyperexponential		Balking + Exponential		Weibull		Erlang	
Data set	MSE	p-value	MSE	p-value	MSE	p-value	MSE	p-value
1	7.13e-5	0.747	6.72e-4	2e-12	7.68e-4	0.002	0.031	1e-30
2	2.32e-4	0.018	8.10e-3	4e-43	2.38e-3	1e-11	0.052	5e-35
3	1.40e-4	0.006	7.05e-4	8.6e-5	1.88e-4	0.006	0.031	6e-28
4	2.67e-5	0.974	6.98e-5	0.518	1.52e-4	0.424	0.014	9e-13

Table 5.2. Comparison of different patience distributions.

The estimated parameters in Table 5.1 warrant additional attention. Since γ_1 is much higher than γ_2, with probability p a delayed customer will quickly abandon. This is equivalent to the modeling of balking with probability α in Model 1. This also agrees with the fact that γ and γ_2 are close to each other. Furthermore, we see that p is close to α and a bit higher. This is as expected since in Model 2 these very impatient customers do not necessarily abandon immediately. Next, we perform a statistical test to assess the fit of our models.

Earlier research by Baccelli and Hebuterne (1981) and Kort (1983) mentioned that the patience distribution could be Erlang with three phases or Weibull. In Table 5.2 we make a comparison between these distributions, together with the hyperexponential distribution and balking plus exponential, for different statistics. The first statistic is the mean squared error (MSE), which should be as low as possible for a good model. The second statistic is the p-value of the Kolmogorov-Smirnov test (see Massey 1951), which tests the null hypothesis that the empirical distribution and the tested distribution come from the same distribution. Values below the default significance level of 0.05 reject this hypothesis.

From the table it is clear that the statistics support the modeling of customers' patience by the hyperexponential distribution. If we look at the p-values of the Kolmogorov-Smirnov test, we observe that the null hypothesis is actually rejected on data sets 2 and 3 at a significance level of 0.05. However, for a significance level of 0.01, the null hypothesis will not be rejected for data set 2. For the model that includes balking these statistics are a bit misleading, since customers that balk are not always represented with a patience of zero in the data.

In conclusion, we presented two models for the patience based on real call center data. The first model is a simple extension of the Erlang A model by allowing customers to balk. The second model is a slightly more advanced model, where the

patience is modeled by the hyperexponential distribution.

5.4 Analysis of Call Center Metrics

Consider a call center model with a single class of customers and s statistically identical, parallel servers. We assume that arrivals follow a Poisson process with rate λ, and that service times are exponentially distributed with rate μ. The queueing discipline is first-come first-served (FCFS). In addition, we let customers be impatient. As discussed in Section 5.3, we denote by T the random variable measuring patience times, and we consider two different ways to model T.

The performance measures we analyze next are based on the assumption that the system has reached steady state. Note that our model unconditionally reaches steady state for any random variable $T \neq 0$, see Garnett et al. (2002) for further details. Let τ be the acceptable waiting time and a be the threshold of short abandonments. In practice, reasonable values for τ and a are for example 20 and 5 seconds, respectively. For some managers, customers who immediately balk or those who enter the queue and quickly abandon before a are not really considered as unsatisfied. Therefore, such customers may not be accounted for in the service-level metric of the call center.

In Table 5.3, we define seven service levels that are useful in practice. We denoted them by SL_i, for $i = 1, \ldots, 7$. We present them, as is customary in call centers, in terms of the numbers of calls that arrive in a certain time period. Later on, we formulate them in terms of the corresponding random variables. The virtual waiting time is defined as the waiting time of customers assuming that they are not abandoning.

What should be the right metric? SL_1 and SL_4 do not give information about abandonments. SL_5 is hard to understand by managers and is also not directly measurable using historical data. For this reason it is, according to our experience, never used in call centers. However, this service-level definition dominates the Erlang A literature. SL_6 does not make difference between waiting prior to service or to abandonment. SL_7 does not give information about waiting.

SL_2 and SL_3 exclude short abandonments which is a good aspect. The main drawback of these two metrics, similarly to all other metrics that use the parameter τ, is that they do not give any information on how long callers that have exceeded τ still have to wait. They entice managers to give priority to callers who have not yet reached the acceptable waiting time, thereby increasing even more the waiting time

SL_1	$\dfrac{\#\text{ answered} \leq \tau}{\#\text{ offered}}$
SL_2	$\dfrac{\#\text{ answered} \leq \tau}{\#\text{ offered} - \#\text{ short abandonments}}$
SL_3	$\dfrac{\#\text{ answered} \leq \tau}{\#\text{ offered} - \#\text{ abandoned} \leq \tau}$
SL_4	$\dfrac{\#\text{ answered} \leq \tau}{\#\text{ answered}}$
SL_5	$\dfrac{\#\text{ virtually answered} \leq \tau}{\#\text{ offered}}$
SL_6	$\dfrac{\#\text{ sojourn in queue} \leq \tau}{\#\text{ offered}}$
SL_7	$\dfrac{\#\text{ abandoned}}{\#\text{ offered}}$

Table 5.3. Service levels.

of callers that have waited longer than τ. Even though they have perverse effects, these metrics are regularly used in practice. One way to avoid unwanted behavior is to add an objective on the performance of the customers who wait more than τ, or to use a different service-level objective. One possibility is to use the time that waiting exceeds τ. In contrast with the expected waiting time (the average speed of answer) it is sensitive to waiting-time variability. Another intuitive and simple solution is to use FCFS in all cases.

5.4.1 Computation of Service Levels

In this subsection, we derive the expressions for the service levels defined in Table 5.3. Let V_Q be the random variable denoting the virtual waiting time of a tagged, infinitely patient customer. In other words if the tagged customer finds a busy system upon arrival, this customer does not balk, neither abandon while waiting in the queue. Note that "answered" means $V_Q \leq T$ and "abandoned" means $V_Q > T$. Let W_Q be the random variable measuring the sojourn time of a customer in the queue. This sojourn time will end either as a result of an abandonment or a start of service. Thus

$$W_Q = \min\{V_Q, T\}.$$

In what follows, we first give the expressions for the service levels in Table 5.3 as a function of the random variables V_Q, W_Q, and T. These expressions will be used later on to fully characterize the service levels. For an event E, $\mathbb{P}(E)$ is defined as the probability that E occurs. We denote by $\neg E$ the complementary event of E, $\mathbb{P}(\neg E) = 1 - \mathbb{P}(E)$. We can write the first service level as

$$\mathrm{SL}_1 = \mathbb{P}(V_Q \leq \tau, V_Q \leq T). \tag{5.1}$$

The second service level is

$$\mathrm{SL}_2 = \frac{\mathbb{P}(V_Q \leq \tau, V_Q \leq T)}{\mathbb{P}(\neg(T < V_Q, T < a))}.$$

Since the patience of a customer is independent of all other events, we have $\mathbb{P}(T > a, V_Q > a) = \mathbb{P}(T > a)\mathbb{P}(V_Q > a)$. Observing now that

$$\mathbb{P}(\neg(T < V_Q, T < a)) = \mathbb{P}(T > a, V_Q > a) + \mathbb{P}(V_Q \leq a, V_Q \leq T),$$

we obtain

$$\mathrm{SL}_2 = \frac{\mathbb{P}(V_Q \leq \tau, V_Q \leq T)}{\mathbb{P}(T > a)\mathbb{P}(V_Q > a) + \mathbb{P}(V_Q \leq a, V_Q \leq T)}. \tag{5.2}$$

Similarly, SL_3 is given by

$$\mathrm{SL}_3 = \frac{\mathbb{P}(V_Q \leq \tau, V_Q \leq T)}{\mathbb{P}(T > \tau)\mathbb{P}(V_Q > \tau) + \mathbb{P}(V_Q \leq \tau, V_Q \leq T)}. \tag{5.3}$$

We also have

$$\mathrm{SL}_4 = \frac{\mathbb{P}(V_Q \leq \tau, V_Q \leq T)}{\mathbb{P}(V_Q \leq T)}, \tag{5.4}$$

$$\mathrm{SL}_5 = \mathbb{P}(V_Q \leq \tau), \tag{5.5}$$

$$\mathrm{SL}_6 = \mathbb{P}(W_Q \leq \tau), \tag{5.6}$$

and finally

$$\mathrm{SL}_7 = \mathbb{P}(V_Q > T). \tag{5.7}$$

In the next subsection, we explicitly derive the previous expressions for the service levels $\mathrm{SL}_1, \ldots, \mathrm{SL}_7$.

5.4.2 Explicit Expressions for Service Levels

The first service level can be obtained from Equation (5.1) and the building blocks of Zeltyn and Mandelbaum (2005) (see Chapter 1) in the following way.

$$\begin{aligned}\mathrm{SL}_1 &= \mathbb{P}(V_Q = 0) + \int_0^\tau \bar{G}(x)v(x)\mathrm{d}x \\ &= \frac{\mathcal{E}}{\mathcal{E}+\lambda J} + \int_0^\tau \bar{G}(x)\frac{\lambda e^{\lambda H(x)-s\mu x}}{\mathcal{E}+\lambda J}\mathrm{d}x.\end{aligned}$$

From

$$\int_0^\tau (\lambda\bar{G}(x) - s\mu)e^{\lambda H(x)-s\mu x}\mathrm{d}x = e^{\lambda H(x)-s\mu x}\Big|_0^\tau = e^{\lambda H(\tau)-s\mu\tau} - 1,$$

it follows that

$$\begin{aligned}\int_0^\tau \lambda\bar{G}(x)e^{\lambda H(x)-s\mu x}\mathrm{d}x &= \int_0^\tau (\lambda\bar{G}(x) - s\mu)e^{\lambda H(x)-s\mu x}\mathrm{d}x \\ &\quad + \int_0^\tau s\mu e^{\lambda H(x)-s\mu x}\mathrm{d}x \\ &= e^{\lambda H(\tau)-s\mu\tau} - 1 + s\mu(J - J(\tau)),\end{aligned}$$

and hence

$$\mathrm{SL}_1 = \frac{\mathcal{E} + e^{\lambda H(\tau)-s\mu\tau} - 1 + s\mu(J - J(\tau))}{\mathcal{E}+\lambda J}.$$

Using that $\mathbb{P}(T > a) = \bar{G}(a)$ and

$$\mathbb{P}(V_Q > a) = \frac{\lambda J(a)}{\mathcal{E}+\lambda J},$$

Equation (5.2) gives

$$\mathrm{SL}_2 = \frac{\mathcal{E} + e^{\lambda H(\tau)-s\mu\tau} - 1 + s\mu(J - J(\tau))}{\bar{G}(a)\lambda J(a) + \mathcal{E} + e^{\lambda H(a)-s\mu a} - 1 + s\mu(J - J(a))}.$$

Similarly, Equation (5.3) becomes

$$\mathrm{SL}_3 = \frac{\mathcal{E} + e^{\lambda H(\tau)-s\mu\tau} - 1 + s\mu(J - J(\tau))}{\bar{G}(\tau)\lambda J(\tau) + \mathcal{E} + e^{\lambda H(\tau)-s\mu\tau} - 1 + s\mu(J - J(\tau))}.$$

Using that $\mathbb{P}(V_Q < T)$ is the probability of service, Equation (5.4) leads to

$$\mathrm{SL}_4 = \frac{\mathcal{E} + e^{\lambda H(\tau) - s\mu\tau} - 1 + s\mu(J - J(\tau))}{\mathcal{E} + s\mu J - 1}.$$

Finally, Equations (5.5)–(5.7) are

$$\mathrm{SL}_5 = 1 - \frac{\lambda J(\tau)}{\mathcal{E} + \lambda J}.$$

$$\mathrm{SL}_6 = 1 - \frac{\lambda \bar{G}(\tau) J(\tau)}{\mathcal{E} + \lambda J}.$$

$$\mathrm{SL}_7 = \frac{1 + (\lambda - s\mu)J}{\mathcal{E} + \lambda J}.$$

Using the specific functional form of the patience distribution T, we can determine $\bar{G}(x)$ and $H(x)$ in closed form. For Model 1, with exponential patience times, these are

$$\bar{G}(x) = (1 - \alpha)e^{-\gamma x},$$
$$H(x) = \frac{1 - \alpha}{\gamma}(1 - e^{-\gamma x}),$$

and for Model 2, with hyperexponential patience times, we get

$$\bar{G}(x) = pe^{-\gamma_1 x} + (1 - p)e^{-\gamma_2 x},$$
$$H(x) = \frac{p}{\gamma_1}(1 - e^{-\gamma_1 x}) + \frac{1 - p}{\gamma_2}(1 - e^{-\gamma_2 x}).$$

The function $J(t)$ cannot be given in closed form for these models. On the other hand, there are no difficulties in evaluating $J(t)$ numerically. We have computed all the expressions for the service levels. These expressions will be used next for the numerical illustrations.

5.5 Numerical Experiments

In this section we first validate the modeling approaches by comparing results of numerical experiments with the empirical results. After that, we show the effect of different service levels on the staffing levels.

5.5.1 Comparison Between the Models

To illustrate the performance of the two models, we consider the four data sets again and compute the service level given by SL_1. The parameters related to the patience distributions of Model 1 and Model 2 are given in Table 5.1. We compare the service-level estimates of these models with the empirical service level. The empirical service level is obtained from the model that directly uses the empirical patience distribution. The analysis is in line with the analysis in Section 5.4, except that the function $H(x)$ has to be evaluated numerically as well.

As a first example, we consider a relatively small system with the following parameters: $\mu = 0.2$, $s = 19$, $\tau = 1/3$, and a varying λ. The empirical service level and the service-level estimates of both models are depicted in Figure 5.2. All plots show that both models have an excellent performance. There are only some small differences noticeable on data sets 2 and 3 for Model 1. This gives us confidence in the usefulness of our models.

As a second example, we consider a larger system defined by $\mu = 0.2$, $s = 210$, $\tau = 1/3$, and a varying λ. The results of the comparison are shown in Figure 5.3. Here we observe mixed results. Model 2 has perfect accuracy on data sets 1 and 4, and a very good performance on data set 2. The service level is underestimated on data set 3. This could perhaps be explained by Figure 5.1, since the fit of the hyperexponential distribution is there not perfect as well. The results for Model 1 are different. The performance on data set 4 is accurate, but the performance on the other data sets is poor. The service level is clearly overestimated. The behavior on data set 2 is also strange.

We also considered SL_7, the abandonment probability. For the sake of conciseness, we only note that both models perform very well with respect to this metric. Only in case of the larger system Model 1 has some minor discrepancies.

All in all, we can conclude from these experiments that both models are useful to model customers' patience for relatively small systems. When the size of the system increases, the model where the patience is modeled by the hyperexponential distribution is preferred. The accuracy of this model compared with empirical results is almost perfect.

5.5.2 Comparison Between the Metrics

Let us consider the metrics SL_1, SL_2, and SL_3. We choose an acceptable waiting time $\tau = 1/3$ minute. The patience is modeled by the empirical patience of

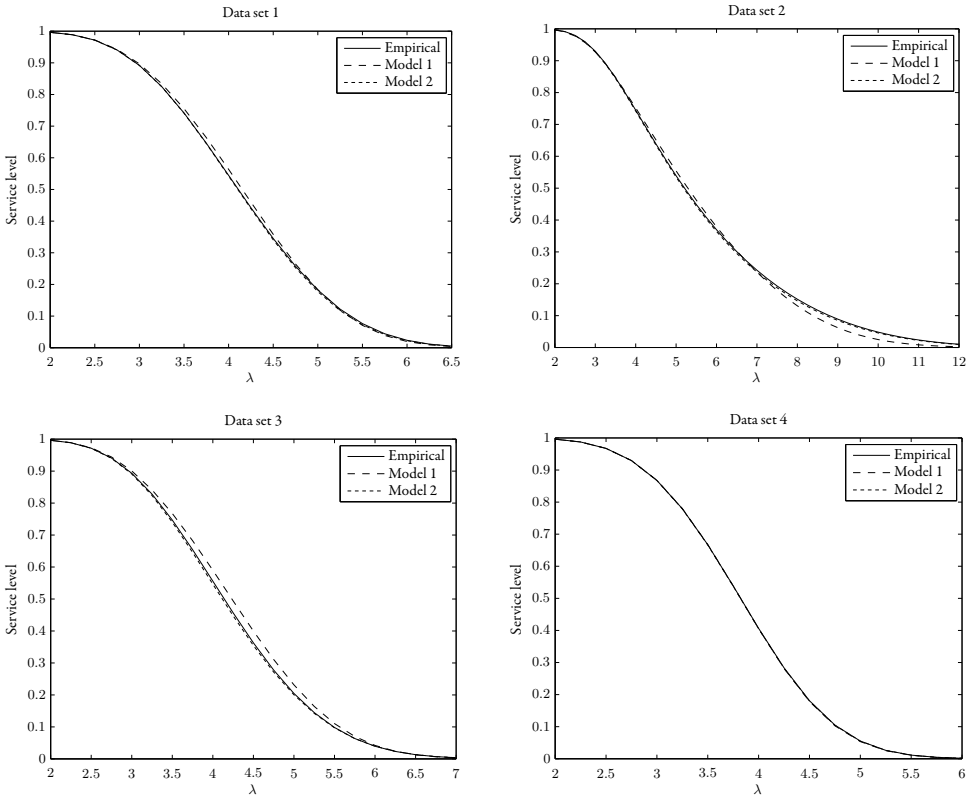


Figure 5.2. Comparison between the models for a small system.

data set 2. Customers who abandon before $a = 5$ seconds are considered as short abandonments, i.e., these are not big issues for the call center manager. The service rate is $\mu = 1$ per minute. We consider different objective values for $\mathrm{SL}_i^* \in \{50\%, 80\%, 95\%, 99\%\}$, for $i = 1, 2, 3$. For each objective SL_i^* and for each $\lambda \in \{3, 5, 7, 10, 15, 20, 30, 50\}$ we compute the optimal staffing level s_i^*. The results are given in Figure 5.4.

Note that in practice managers usually use SL_1 which is not appropriate since we are penalized with customers who are very impatient. These customers do not really experience frustration. A better metric would be SL_2 which ignores short abandonments. An even better metric could be SL_3 which ignores abandonments within the acceptable waiting time. An additional benefit from using these last two metrics, as shown in Figure 5.4, is that the required staffing levels are lower

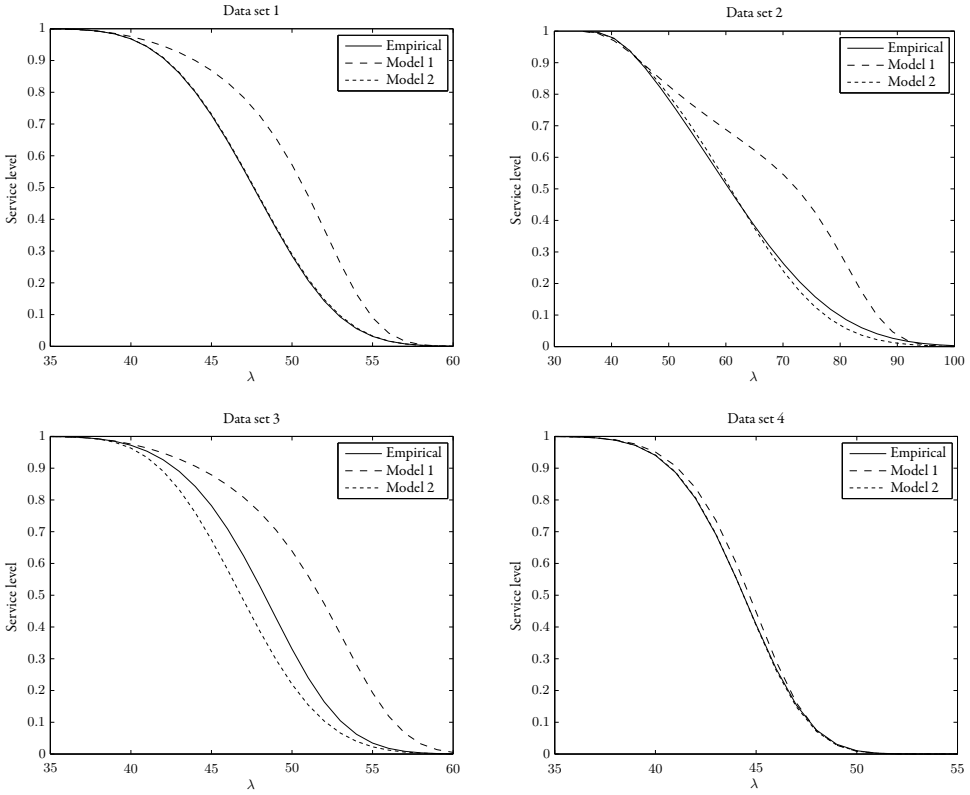


Figure 5.3. Comparison between the models for a large system.

than that for SL_1.

To go further and confirm the interest of SL_2 and SL_3, it is worth to look on the behavior of the probability of abandonment. To do so, we consider the case $\mathrm{SL}_i^* = 80\%$ with the same optimal staffing levels s_i^* as shown in Figure 5.4 ($i = 1, 2, 3$). In Table 5.4, we vary λ and give both the probabilities of abandonment, SL_7, and that of abandoning after τ, say SL_8. The latter can been seen as a "reasonable" probability of abandonment, which does not include customers who abandon before τ. SL_8 can be computed as follows. From Equations (5.1) and (5.3), we may write

$$\mathbb{P}(\text{abandon before } \tau) = 1 - \frac{\mathrm{SL}_1}{\mathrm{SL}_3}. \tag{5.8}$$

Knowing that $\mathrm{SL}_7 = \mathbb{P}(\text{abandon before } \tau) + \mathbb{P}(\text{abandon after } \tau)$, Equation (5.8)

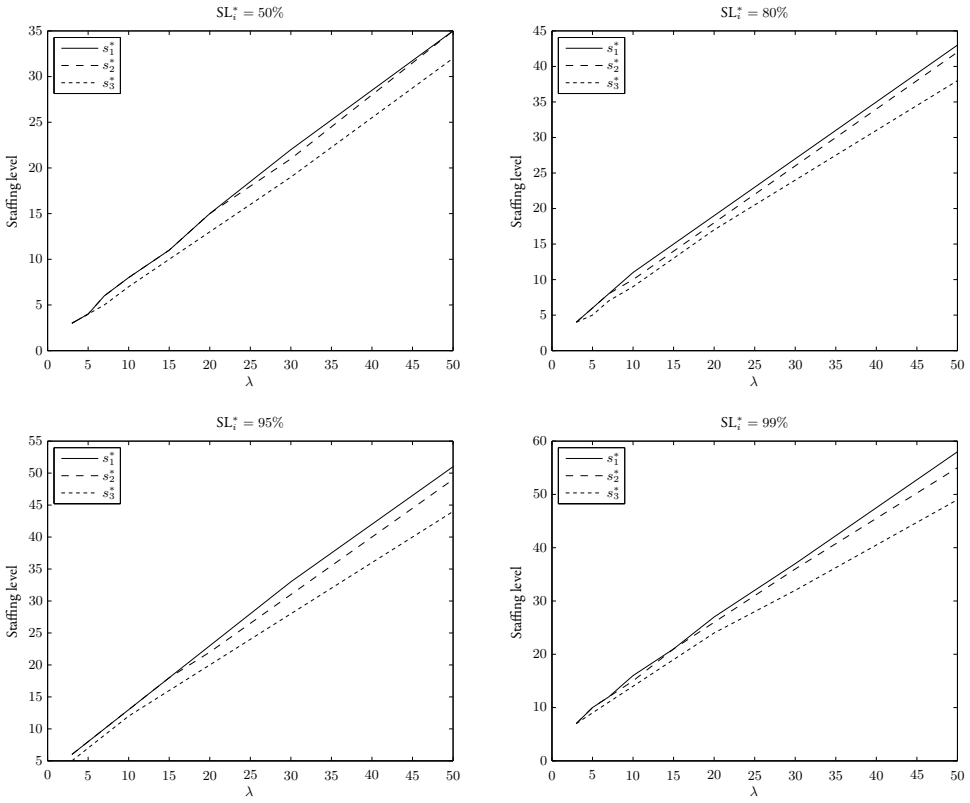


Figure 5.4. Optimal staffing levels.

leads to

$$\mathrm{SL}_8 = \mathrm{SL}_7 + \frac{\mathrm{SL}_1}{\mathrm{SL}_3} - 1.$$

From Table 5.4, we see that the performance in terms of abandonments after τ are acceptable for the metrics SL_2 and SL_3 (while they do need lower staffing levels). This comment is particularly relevant for large call centers, due to the benefit of pooling on performance.

5.5.3 Extension to the Case with Retrials

In this subsection, we extend the analysis by allowing retrials. In practice, some of the customers who balk or abandon will redial and try to access the call center

λ	3	5	7	10	15	20	30	50
s_1^*	4	6	8	11	15	19	27	43
s_2^*	4	6	8	10	14	18	26	42
s_3^*	4	5	7	9	13	17	24	38
SL_7 with s_1^*	0.112	0.104	0.096	0.086	0.105	0.117	0.133	0.151
SL_7 with s_2^*	0.112	0.104	0.096	0.129	0.141	0.148	0.158	0.168
SL_7 with s_3^*	0.112	0.182	0.155	0.184	0.184	0.184	0.212	0.242
SL_8 with s_1^*	0.022	0.016	0.012	0.008	0.007	0.006	0.005	0.003
SL_8 with s_2^*	0.022	0.016	0.012	0.014	0.012	0.010	0.007	0.003
SL_8 with s_3^*	0.022	0.035	0.023	0.025	0.019	0.015	0.013	0.010

Table 5.4. Probability of abandonment (SL_7), and probability of abandoning after τ (SL_8) for $SL_i^* = 80\%$.

again. For more details on the modeling and analysis of call centers with retrials, we refer the reader to Aguir et al. (2004) and Pustova (2010), and references therein.

In what follows, we consider a simple modeling of the retrials and analyze its impact on the optimal staffing level. We want to assess how ignoring retrials may lead to an insufficient staffing level. Let us consider a model with generally distributed patience times. We allow some of the customers who balk or abandon to call back the call center. We denote by θ the probability that one will call back. Delays before customers' call back are assumed to be i.i.d. random variables with a general distribution. For tractability purposes, we assume independence between successive calls in terms the probability to call back. Let $\bar{\lambda}$ be the overall arrival rate to the system, i.e., the sum of the primary calls and the feedback calls. Then,

$$\bar{\lambda} = \frac{\lambda}{1 - \theta \mathbb{P}(A \mid \bar{\lambda})}, \tag{5.9}$$

where λ is the arrival rate of the primary calls, and the probability to abandon $\mathbb{P}(A \mid \bar{\lambda}) = SL_7$ is computed based on $\bar{\lambda}$. By varying θ from 0 to 1, we move from our original system with no retrials to a system with high retrials. This simple model falls into the class of product-form networks analyzed by Baskett et al. (1975). As a result, the stationary behavior of this queueing model does not depend on the distribution of the call-back delays. They can thus be ignored. Using this simple modeling, we capture the retrial feature while being able to use the results developed

in Section 5.4. To evaluate the performance of a call center with parameters λ, μ, s, θ, and generally distributed patience times, it suffices to use the results of Section 5.4 for a call center with parameters $\bar{\lambda}$, μ, s, and the same patience distribution.

Before going further, we should discuss how to compute $\bar{\lambda}$. Denoting the right-hand side in Equation (5.9) by a continuous function g in $\bar{\lambda}$, we may write $\bar{\lambda} = g(\bar{\lambda})$. Then, $\bar{\lambda}$ is said to be a fixed point of g. To numerically compute $\bar{\lambda}$, we propose the following fixed-point algorithm.

FIXED-POINT ALGORITHM()

Initialization: $\bar{\lambda}_0 \leftarrow \lambda$, $i \leftarrow 0$, ϵ

Do

$i \leftarrow i + 1$

$\bar{\lambda}_i \leftarrow \lambda + \bar{\lambda}_{i-1}\theta\mathbb{P}(A \mid \bar{\lambda}_{i-1})$

While $\left|\frac{\bar{\lambda}_i - \bar{\lambda}_{i-1}}{\bar{\lambda}_{i-1}}\right| > \epsilon$

$\bar{\lambda} \leftarrow \bar{\lambda}_i$

END ALGORITHM

In what follows, we want to numerically study the impact of retrials on the optimal staffing level. We choose the metric SL_2, and our objective is to compute the optimal staffing level for $\mathrm{SL}_2^* = 80\%$, $a = 5$ seconds, $\tau = 1/3$ minute. We consider different call center sizes. The primary arrival rates are $\lambda \in \{10, 30, 50\}$. We also consider different levels of retrials, $\theta \in [0, 1]$. Similarly to the previous subsection, we choose $\mu = 1$ and model the patience times by the empirical distribution of data set 2. The results are given in Figure 5.5. As expected, this figure reveals that the optimal staffing level, s_2^*, increases in the probability to call back θ. An important observation here is that the impact of call backs on s_2^* can be high, although we start from a model with no retrials ($\theta = 0$) that has a high service level ($\mathrm{SL}_2 = 80\%$). For instance, the optimal staffing level can increase by about 20% when moving from a system with no retrials to a system with high retrials ($\theta = 1$). This is particularly due to the high impatience of customers in data set 2. In such cases, ignoring the modeling of retrials may lead to inappropriate results.

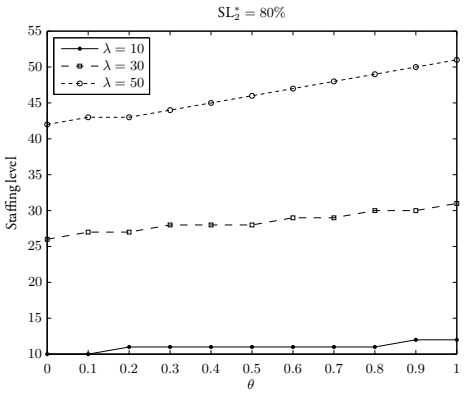


Figure 5.5. The impact of retrials.

5.6 Conclusion

We have analyzed various process-related call center metrics that include customer abandonment. We showed how to obtain existing results explicitly, and also derived new results for new metrics considering short abandonments or abandonments within the acceptable waiting time. In practice, many managers choose not to count short abandonments against the call center performance metrics. Although the models used here are simple (with Markovian assumptions), we have shown their robustness using real call center data. Through numerical analysis, we have also discussed the advantages and disadvantages of the different metrics.

We have presented two models for customers' patience that have a very good agreement with reality. The method to derive the call center metrics works for empirical patience distributions as well. The benefit of using our models is that the Markovian property is preserved. This is especially useful when one wants to consider other service-time distributions.

There are several avenues for future research. It would be useful to extend the analysis to the case of more than one customer type with nonidentically distributed patience and service times. Another interesting and challenging extension of the current analysis is to consider a nonstationary arrival process.

Chapter 6

Queueing Delays of Priority Queues with Impatience

In the previous chapter, we extensively analyzed Markovian queueing systems with impatient customers. In this chapter, we additionally consider the feature of priorities, where high-priority customers get nonpreemptive priority over the low-priority ones. For each type of customers, we focus on various performance measures related to queueing delays. We consider two cases where the discipline of service within each customer type is FCFS or LCFS. We explicitly derive the Laplace-Stieltjes transforms of the random variables. Finally we compare FCFS and LCFS and gain insights through numerical experiments.

6.1 Introduction

In this chapter, we analyze queueing systems with multiple types of impatient customers. Customer abandonment (or also reneging) is an important feature in a wide variety of situations that may be encountered in telecommunication systems, manufacturing systems, and service systems such as call centers and health care systems. Theoretical models incorporating abandonment are therefore closer to reality and necessary to obtain more accurate analysis. Another important feature in practice is the differentiation in the service given to different customer types. A priority mechanism is a useful scheduling method that allows different customer types to receive differentiated performance levels. Priority queueing comes up in many applications such as communication networks with differentiated services, call centers with VIP and less important customers, and more. Priority schemes are additionally known for their ease of implementation, explaining their prevalence in practice. Much of the queueing literature is devoted to analyzing priority queues. Most papers are restricted to two priority types. There are two possible refinements

in priority situations, namely preemption and nonpreemption. In the preemptive case, a customer with high priority is allowed to enter service immediately even if another one with lower priority is already present in service. On the other hand, a priority discipline is said to be nonpreemptive if there is no interruption. A customer with higher priority just goes to the head of the queue and waits for his or her turn.

We consider a Markovian multi-server queueing system with two types of impatient customers: high- and low-priority ones. The high-priority type has nonpreemptive priority over the other type. We assume common exponential distributions for service times as well as patience times for both customer types. We analyze two different systems by considering different disciplines of service within each queue. The discipline of service of a queue refers to the manner by which customers are selected for service when a queue has formed. The most common discipline that can be observed in everyday life is first-come first-served (FCFS). Some other in common usage are random order of service (ROS) and last-come first-served (LCFS), which is applicable to many inventory systems when it is easier to reach the nearest stored items which are the last in. In this chapter, we consider FCFS and LCFS policies and derive various performance measures related to queueing delays. Our approach is based on the use of Laplace-Stieltjes transforms and on the characterization of the virtual waiting time of a "virtual" infinitely patient customer. We also describe the procedure to extend the analysis to more than two customer types.

Our motivation for considering identical statistical behavior of customer types (service and patience times) relates to the type of models that motivate our analysis. We are considering firms where customers are segmented into different groups based on their value to the firm. This segmentation can be based on lifetime value or profitability. The company then provides different levels of service to these groups. This type of service-level differentiation is widely used in financial service, telecommunication call centers, and more. In the presence of this type of segmentation, the difference between customer types is not related to the statistical behavior of customers but to their importance for the company, which we capture through priorities. In concrete terms, we assume for our models that customer behavior and queries do not differ from one type to another. This is a reasonable assumption for such systems, see Zeltyn et al. (2009).

In what follows, we review some of the queueing literature related to this chapter. We distinguish two streams of literature. The first deals with queueing models

with impatient customers. The second focuses on priority queues. The literature on queueing models with abandonments focuses especially on performance evaluation. We refer the reader to Chapter 5 for an overview on this literature.

Let us now briefly mention some of the literature dealing with priority queueing systems. We refer the reader to Davis (1966) and Kella and Yechiali (1985) for a simple Markovian nonpreemptive queue where all customer types have the same service-time distribution. Wagner (1997) considers multi-server nonpreemptive priority systems with a Markovian arrival process, service times having phase type distributions, and both cases of finite and infinite queueing spaces are considered. Other references considering more complicated models, but where abandonments are not allowed, include those by Kao and Wilson (1999), Takine (1999), and Sleptchenko (2003). As for preemption schemes, we refer the reader to Harchol-Balter et al. (2005), Sleptchenko and Van der Heijden (2005), and references therein. Sleptchenko and Van der Heijden (2005) derive approximations for a wide range of relevant performance characteristics, such as the moments of the number of customers of a certain type, in a Markovian queue where customers have different expected values of service times. Harchol-Balter et al. (2005) introduce a new technique to reduce the Markov chain dimensionality of an $M/PH/s$ model with an arbitrary number of preemptive-resume priority types. Some research on priority queues has been dedicated to systems with mixed priorities that combine the two disciplines (with and without preemption). Results for the single-server case can be found in Drekic and Stanford (2000), and for those in the multi-server case, we refer the reader to Zeltyn et al. (2009).

Although the two features of abandonment and priority have each received attention separately, there is limited literature that deals with both of them. We refer the reader to Choi et al. (2001), where the authors derive several performance measures for an $M/M/1$ queue with two types of impatient customers in which type 1 customers have impatience of constant duration, and type 2 customers have no impatience and low priority level. An extension of the latter model is addressed by Brandt and Brandt (2004) for general distributed patience times. Rozenshmidt (2007) considers a similar model to ours (under FCFS) and derives expressions for the unconditional expected waiting times of all customer types. Here we extend that analysis by considering additional performance measures, by considering also LCFS, and by computing all moments of the random variables. We also refer the reader to an interesting paper by Iravani and Balcıoğlu (2008b), where the authors analyze different priority models: single-server models with general service times,

and multi-server models with exponential service times and a call-back option. Our main contributions can be summarized as follows.

- We compute the Laplace-Stieltjes transforms of various random variables related to queueing delays: unconditional waiting time, and conditional waiting times given service and given abandonment. We do so for both high- and low-priority customers. Our approach is based on the computation of virtual waiting times. One can then easily numerically invert the Laplace-Stieltjes transforms in order to obtain the cumulative distribution functions of these random variables at any point of time, see Abate and Whitt (2006).

- The analysis is detailed for two different nonpreemptive priority models. One where the discipline of service within each class is FCFS, and another one working under LCFS. Moreover, the analysis we develop holds for a priority queue with mixed policies, i.e., FCFS for the first type and LCFS for the second one, and vice versa. We also extend our approach to the case of more than two customer types.

- We numerically compare the performance measures of the FCFS and LCFS policies, and provide some managerial insights.

The remainder of this chapter is structured as follows. In Subsections 6.2.1 and 6.2.2, we describe the basic two-class queueing models, and define the performance measures of interest, respectively. In Subsection 6.2.3, we then develop some preliminary results that would help us in the rest of the analysis. In Subsection 6.3.1, we provide the results of performance evaluation when high- and low-priority customers are served under the FCFS basis. Those when high- and low-priority customers are served under the LCFS basis are given in Subsection 6.3.2. In Subsection 6.3.3, we explain how the analysis can be extended to the case of more than two classes. In order to illustrate the results and compare FCFS and LCFS, we give some numerical experiments in Subsection 6.3.4. Finally in Section 6.4, we provide some concluding remarks and directions for future research.

6.2 Preliminaries

We first describe the two basic multi-class queues (for FCFS and LCFS) that we will analyze in this chapter. Then, we provide the definitions of the performance

measures of interest. The performance measures are related to the queueing delays of customers. Finally, we present some preliminary derivations that we will need along the way.

6.2.1 Modeling

Consider a queueing model with two types of customers: important customers denoted by type 1, and less important ones denoted by type 2. The model consists of two infinite-buffer queues for types 1 and 2, and a set of s parallel, identical servers. All servers are able to handle all types of customers. The system is work conserving, i.e., a server is never forced to be idle with customers waiting. So upon arrival, a customer is addressed by one of the available servers, if any. If not, the customer must join one of the queues. Newly arriving customers of types 1 and 2 are assigned to queues 1 and 2, respectively. Customers of type 1 (waiting in queue 1) have priority over customers of type 2 (waiting in queue 2) in the sense that agents are providing assistance to type 1 customers first. The priority rule is nonpreemptive, which simply means that a server currently serving a type 2 customer, while a new type 1 customer enters the system, will complete this service before turning to the queue 1 customer. Within each queue, we consider two cases for the discipline of service: FCFS and LCFS. Arrival processes of types 1 and 2 follow a Poisson process with rates λ_1 and λ_2, respectively. Let λ be the total arrival rate, $\lambda = \lambda_1 + \lambda_2$. Successive service times are assumed to be independent and identically distributed (i.i.d.), and follow a common exponential distribution with rate μ for both customer types.

In addition, we let customers be impatient. After entering the queue, a customer will wait a random length of time for service to begin. If service has not begun by this time the customer will abandon. Times before abandonment, for both customer types, are assumed to be i.i.d. and exponentially distributed with a common rate denoted by γ. We describe patience times by the random variable T. Finally, retrials are ignored, and abandonment is not allowed once a customer starts service. Following similar arguments, the behavior of the system can be viewed as a two-class $M/M/s+M$ queueing system. The resulting model where the policy for each queue is FCFS (LCFS) is referred to as $\text{Model}_{\text{FCFS}}$ ($\text{Model}_{\text{LCFS}}$). Note that owing to abandonments, $\text{Model}_{\text{FCFS}}$ and $\text{Model}_{\text{LCFS}}$ are unconditionally ergodic.

6.2.2 Notation

We denote by m the type of a customer, $m \in \{1, 2\}$. During the stationary regime, we define the following performance measures for $\text{Model}_{\text{FCFS}}$ and $\text{Model}_{\text{LCFS}}$. To simplify the notations, we will not add indices to these quantities in order to refer to one of the models (we will add a clarification comment when it is necessary). In the remainder of this chapter we refer to a customer as a she.

- W is the unconditional queueing delay of an arbitrary customer (regardless of her type).
- W_m is the unconditional queueing delay of a type m customer.
- $W_{m,s}$ is the conditional queueing delay of a type m customer, given that she will enter service.
- $P_{m,s}$ is the probability that a type m customer enters service.
- $W_{m,r}$ is the conditional queueing delay of a type m customer, given that she will abandon.
- $P_{m,r}$ is the probability that a type m customer abandons.
- $W_{m,d}$ is the conditional queueing delay of a type m customer, given that she has to wait.
- P_d is the probability of delay, i.e., the probability that a new arrival has to wait. Since $\text{Model}_{\text{FCFS}}$ and $\text{Model}_{\text{LCFS}}$ are work conserving, P_d is independent of the customer type.
- $W_{m,d,s}$ is the conditional queueing delay of a type m customer, given that she was queued and that she will enter service. (We do not define a similar quantity for abandoned customers, since an abandoned customer is necessarily a delayed customer.)
- $P_{m,d,s}$ is the probability that a type m customer waiting in the queue will enter service.

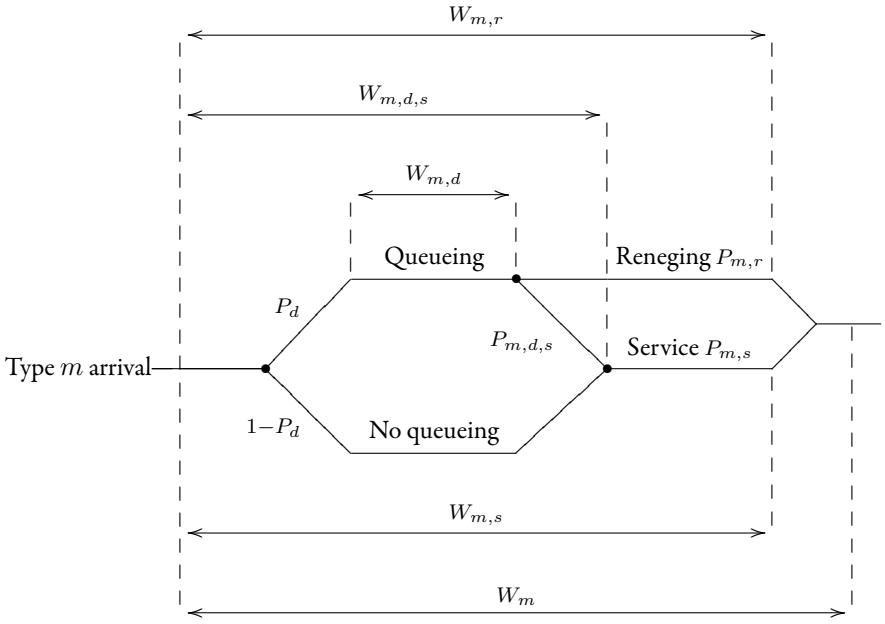


Figure 6.1. Performance measures for a type m customer.

To clarify the numerous definitions, we depicted in Figure 6.1 a schema of the performance measures of interest.

In what follows, we provide some relations between the performance measures. For the remainder of the chapter, we denote by $\mathbb{E}X^k$ the k-th order moment of a given random variable X, for $k \geq 1$. We also denote by $f_X(\cdot)$ and $F_X(\cdot)$ the probability density function (pdf) and the cumulative distribution function (cdf) of X. A customer who does not abandon will necessarily enter service, then $P_{m,s} + P_{m,r} = 1$. A customer who joins the queue has two possibilities: either she abandons, or she gets service, so $P_d = P_{m,r} + P_{m,d,s}$. Since the arrival processes are Poisson, the probability that a new arrival is of type m is λ_m/λ. Therefore,

$$\mathbb{E}W^k = \frac{\lambda_1}{\lambda}\mathbb{E}W_1^k + \frac{\lambda_2}{\lambda}\mathbb{E}W_2^k,$$

for $k \geq 1$. For type m customers, one may write

$$\mathbb{E}W_m^k = P_{m,s}\mathbb{E}W_{m,s}^k + P_{m,r}\mathbb{E}W_{m,r}^k, \tag{6.1}$$

for $k \geq 1$. Upon arrival, a customer is immediately addressed by one of the available servers, if any. If not, she has to wait and joins one of the queues (with probability P_d). Thus,

$$\mathbb{E}W_m^k = P_d \mathbb{E}W_{m,d}^k, \tag{6.2}$$

for $k \geq 1$. For customers that join the queue, we have

$$\mathbb{E}W_{m,d}^k = P_{m,d,s} \mathbb{E}W_{m,d,s}^k + P_{m,r} \mathbb{E}W_{m,r}^k, \tag{6.3}$$

which allows to determine $\mathbb{E}W_{m,d,s}^k$, for $k \geq 1$.

6.2.3 Preliminary Analysis

We start by computing the stationary probability distributions of the system states for $\text{Model}_{\text{FCFS}}$ and $\text{Model}_{\text{LCFS}}$. At a given instant t, we denote by $n_1(t)$, $n_2(t)$, and $n(t) = n_1(t) + n_2(t)$ the number of type 1 customers in queue 1, that of type 2 in queue 2, and the total in both queues, respectively. Computing the stationary distribution of the process $\{n_2(t), t \geq 0\}$ or $\{(n_1(t), n_2(t)), t \geq 0\}$ is a complicated task. We only consider the processes $\{n_1(t), t \geq 0\}$ and $\{n(t), t \geq 0\}$ which are sufficient for the derivation of the performance measures. Recall that all stationary probabilities exist due to the ergodicity condition (which holds for any $\gamma > 0$).

Patience times are memoryless. Thus, as long as the scheduling policy within each queue is work conserving, the number of type 1 customers and type 2 customers in the system remain unchanged. Moreover, since patience as well as service times are identically distributed for both customer types, a work-conserving policy (priority between the queues or not) does not affect the total number of customers in the system. The following analysis holds for both $\text{Model}_{\text{FCFS}}$ and $\text{Model}_{\text{LCFS}}$.

Let us consider the process $\{n(t), t \geq 0\}$. With regard to the total number of customers in the system ($\text{Model}_{\text{FCFS}}$ or $\text{Model}_{\text{LCFS}}$), our system is equivalent to a multi-server queue with a single type of customers. The arrival process is Poisson with intensity $\lambda = \lambda_1 + \lambda_2$. Hence, this system corresponds to the system analyzed in Section 1.3. The stationary probability distribution of i customers in the system,

denoted by π_i for $i \geq 0$, is given by

$$\pi_i = \begin{cases} \dfrac{\lambda^i}{\mu^i i!}\pi_0, & 0 \leq i \leq s, \\ \dfrac{\lambda^i}{\mu^s s! \prod_{j=1}^{i-s}(s\mu + j\gamma)}\pi_0, & i > s, \end{cases}$$

with

$$\pi_0^{-1} = \sum_{i=0}^{s} \frac{\lambda^i}{\mu^i i!} + \sum_{i=s+1}^{\infty} \frac{\lambda^i}{\mu^s s! \prod_{j=1}^{i-s}(s\mu + j\gamma)}.$$

Denote by $p(i)$ the stationary probability that all servers are busy and there are i customers in total in both queues, i.e., $p(i) = \pi_{s+i}$, for $i \geq 0$.

The probability of delay P_d is simply the probability that a new arrival finds all servers busy. It is then independent of the type of the new arrival. Moreover since the arrival process of a type m customer follows a Poison process, we use the PASTA property to state that the stationary probabilities seen by a new arrival coincide with those seen at an arbitrary instant. Thus, P_d is given by

$$P_d = 1 - \sum_{i=0}^{s-1} \pi_i.$$

Let us now characterize the stationary distribution of $\{n_1(t), t \geq 0\}$. To do so, we consider a two-dimensional Markov chain as shown in Figure 6.2. The state of the system is defined by the total number of customers in the system (regardless of their type) if less than s customers are in the system (i.e., all customers are in service), and defined by the couple (n_1, n_2) denoting the number of queued customers of each type if s customers or more are in the system (i.e., all servers are busy). Let $p_1(i)$ denote the stationary probability that all servers are busy and i type 1 customers are in queue 1. By assembling all the states of each line in Figure 6.2, the balance equations lead to

$$p_1(i) = \frac{\lambda_1^i}{\prod_{j=1}^{i}(s\mu + j\gamma)} p_1(0), \tag{6.4}$$

for $i \geq 0$. To compute $p_1(0)$, we come back to the process $\{n(t), t \geq 0\}$. It is clear that the probability to be in state i, for $0 \leq i \leq s-1$, in the Markov chain of

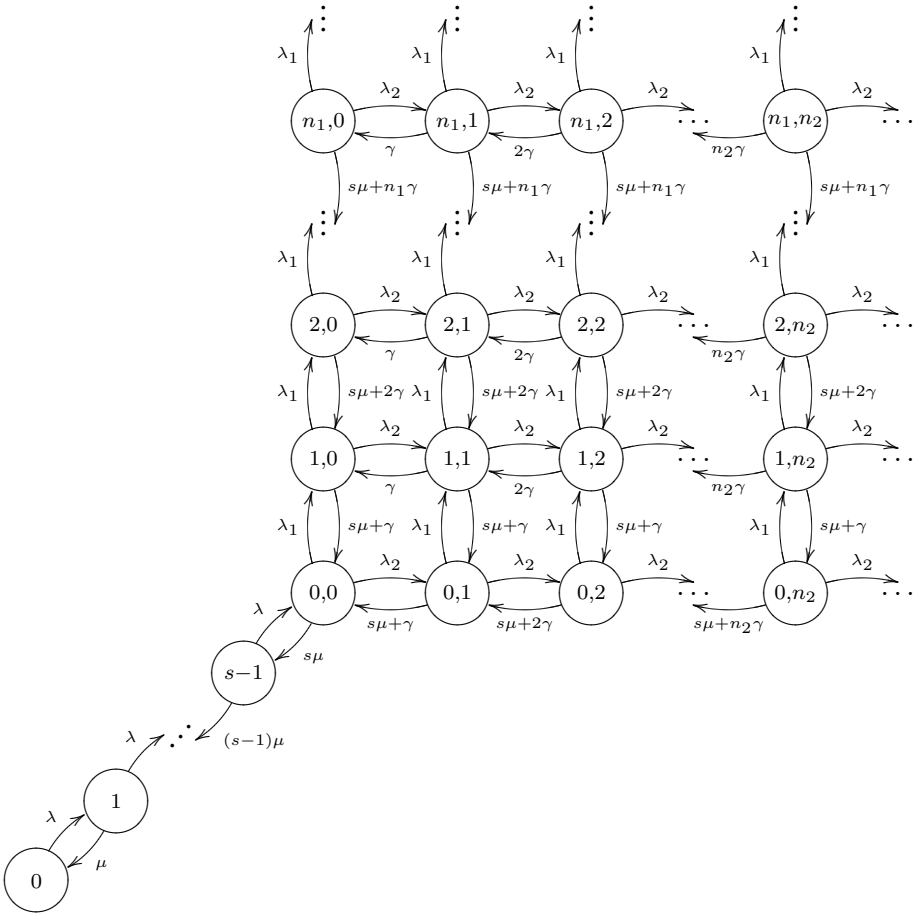


Figure 6.2. Markov chain for the number of customers in the queue.

Figure 6.2 is equivalent to π_i. Next, the normalization condition gives

$$\sum_{i=0}^{s-1} \pi_i + \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} p_{1,2}(i,j) = 1, \tag{6.5}$$

where $p_{1,2}(i,j)$ is the stationary probability that all servers are busy, i type 1 cus-

tomers are in queue 1, and j type 2 customers are in queue 2. Observe now that

$$p_1(i) = \sum_{j=0}^{\infty} p_{1,2}(i,j), \tag{6.6}$$

for $i \geq 0$. Combining thereafter Equations (6.4)–(6.6) leads to

$$p_1(0) = \left(1 - \sum_{i=0}^{s-1} \pi_i\right) \left(\sum_{i=0}^{\infty} \frac{\lambda_1^i}{\prod_{j=1}^{i}(s\mu + j\gamma)}\right)^{-1}.$$

Having in hand $p_1(i)$ and $p(i)$, for $i \geq 0$, the expected length of queue 1, say Q_1, and that of both customer types waiting in both queues, say Q, are therefore given by

$$Q_1 = \sum_{i=1}^{\infty} i p_1(i), \quad \text{and} \quad Q = \sum_{i=1}^{\infty} i p(i). \tag{6.7}$$

As a consequence, the stationary expected length of queue 2, say Q_2, is simply deduced by $Q_2 = Q - Q_1$.

We are now ready to compute the stationary probability to abandon and that to enter service for a new type m arrival. The probability $P_{m,r}$ can be viewed as the fraction of the stationary expected rate of type m abandoned customers over that of type m arrivals, seen at the epoch of a new type m arrival. Using PASTA and the memoryless property of patience times, we deduce that the stationary expected rate of type m abandoned customers is γQ_m. So,

$$P_{m,r} = \frac{\gamma Q_m}{\lambda_m}.$$

The probability to enter service is only the complementary probability (no possible events of blocking or balking). Indeed, a customer who does not abandon will necessarily enter service,

$$P_{m,s} = 1 - P_{m,r}.$$

Finally, we also have

$$P_{m,d,s} = P_d - P_{m,r}.$$

6.3 Analysis of Queueing Delays

In this section, we characterize the distributions of the random variables W_m, $W_{m,d}$, $W_{m,s}$, $W_{m,r}$ and $W_{m,d,s}$. We do so by computing their k-th order moments, for $k \geq 1$. Although the stationary probabilities of the number of type m customers in the system, as well as P_d, $P_{m,s}$, $P_{m,r}$ and $P_{m,d,s}$ (computed in Subsection 6.2.3) are independent of the scheduling policy within each queue, the random variables of queueing delays do depend on the policy (FCFS or LCFS). We separately address the analyses for $\text{Model}_{\text{FCFS}}$ and $\text{Model}_{\text{LCFS}}$ in Subsections 6.3.1 and 6.3.2, respectively.

Our approach is based on the computation of first-passage times in various birth-death processes. As we will prove below, many of these random variables are equivalent to the length of an n-busy period in an FCFS $M/M/s+M$ queue, for $n \geq 0$. For $n \geq 1$, an n-busy period is defined as the elapsed time from the arrival of a customer to a busy $M/M/s+M$ system with $n-1$ waiting customers in the queue (n customers in the queue including the new arrival) until the epoch at which one server becomes idle. The 0-busy period reduces to the classical busy-period definition defined to begin with the arrival of a customer to a system with $s-1$ busy servers and to end when again one server becomes idle. We denote the length of an n-busy period by $BP_{n,\lambda}$, for $n \geq 0$. For an FCFS $M/M/1+M$ queue, one can obtain from Rao (1967) or Iravani and Balcıoğlu (2008b) the Laplace-Stieltjes transform of the pdf of $BP_{n,\lambda}$. Next, using Jouini (2012, Lemma 1) to state that the busy-period distribution is unchanged for all work-conserving policies, substituting the expected service rate of a busy $M/M/1+M$ queue, μ, by that of an $M/M/s+M$ queue, $s\mu$, and denoting the Laplace-Stieltjes transform of the pdf of $BP_{n,\lambda}$ (for an $M/M/s+M$ queue with any work-conserving policy) by $\widetilde{F}_{BP_{n,\lambda}}(x)$, we get

$$\widetilde{F}_{BP_{n,\lambda}}(x) = \frac{\frac{s\mu}{x+s\mu} + \sum_{i=1}^{\infty}(-1)^i \left[\prod_{j=0}^{i-1}\left(1-\frac{s\mu}{x+s\mu+j\gamma}\right)\right]\frac{s\mu}{x+s\mu+i\gamma}\Theta(n,i)}{1+\sum_{i=1}^{\infty}\frac{\lambda^i}{i!\gamma^i}\left[\prod_{j=0}^{i-1}\left(1-\frac{s\mu}{x+s\mu+j\gamma}\right)\right]}, \tag{6.8}$$

with

$$\Theta(n,i) = \begin{cases} \displaystyle\sum_{j=0}^{i} \frac{(-1)^j \lambda^j}{j!\gamma^j} \binom{n}{i-j}, & 1 \leq i \leq n, \\ \displaystyle\sum_{j=i-n}^{i} \frac{(-1)^j \lambda^j}{j!\gamma^j} \binom{n}{i-j}, & i > n, \end{cases}$$

for $x \in \mathbb{R}^+$, and $n \geq 0$. We will later use Equation (6.8) to analyze queueing delays for low-priority customers in $\text{Model}_{\text{FCFS}}$, and both customer types in $\text{Model}_{\text{LCFS}}$. The analysis for high-priority customers in $\text{Model}_{\text{FCFS}}$ is in turn simpler by extending existing results in the literature.

6.3.1 Analysis of $\text{Model}_{\text{FCFS}}$

For high- and low-priority customers, we compute the k-th order moment of $W_{m,s}$ and $W_{m,r}$, which also allows to derive the k-th order moment of W_m, $W_{m,d}$ and $W_{m,d,s}$, for $k \geq 1$ and $m \in \{1, 2\}$.

High-Priority Customers

Using an approach originally inspired by Whitt (1999), Jouini et al. (2011a) derive all moments of $W_{1,s}$ and $W_{1,r}$ in the case of a finite multi-server queue with a single type of impatient customers. Here we further extend that approach for our priority queue. Consider a new type 1 arrival who finds all servers busy and n_1 waiting customers ahead of her in queue 1, $n_1 \geq 0$. It goes without saying that for the remaining cases (at least one server is idle), our customer will immediately enter service. Because of their lower priority, type 2 customers already waiting in queue 2, as well as those who will arrive later, will not affect the sojourn time in the queue of our new type 1 customer. Using Jouini et al. (2011a), we obtain

$$\mathbb{E}W_{1,s}^k = \frac{1}{P_{1,s}} \sum_{n_1=0}^{\infty} p_1(n_1) \Psi_{n_1+1} \mathbb{E}Y_{n_1+1}^k,$$

with

$$\Psi_{n_1} = \prod_{i=1}^{n_1} \left(1 - \frac{\gamma}{s\mu + i\gamma}\right) = \frac{s\mu}{s\mu + n_1\gamma},$$

for $n_1 \geq 1$, and Y_{n_1}, a random variable, is the summation of n_1 independent exponential distributions with parameters $s\mu + \gamma, s\mu + 2\gamma, \ldots, s\mu + n_1\gamma$. So, all moments of Y_{n_1} may be derived in a closed form. For example, its first two moments are

$$\mathbb{E}Y_{n_1} = \sum_{j=1}^{n_1} \frac{1}{s\mu + j\gamma}$$

and

$$\mathbb{E}Y_{n_1}^2 = \sum_{j=1}^{n_1} \frac{1}{(s\mu + j\gamma)^2} + \left(\sum_{j=1}^{n_1} \frac{1}{s\mu + j\gamma} \right)^2,$$

respectively, for $n_1 \geq 1$

Let us now focus on deriving $\mathbb{E}W_{1,r}^k$. For a new type 1 arrival who finds at least one idle server, $W_{1,r}$ is zero. Assume she is queued with n_1 waiting customers and that she will abandon while waiting in the queue. Let Z_{n_1+1} denote the random variable measuring her sojourn time in the queue before abandonment. Removing the condition on n_1, we obtain

$$\mathbb{E}W_{1,r}^k = \frac{1}{P_{1,r}} \sum_{n_1=0}^{\infty} p_1(n_1) \mathbb{E}Z_{n_1+1}^k.$$

Note that computing the moments of Z_{n_1}, for $n_1 \geq 1$, again involves summations of independent exponential random variables, and are easy to obtain. One may see that the probability to abandon at position j, for $1 \leq j \leq n_1$, is

$$\frac{\gamma}{s\mu + j\gamma} \prod_{l=j+1}^{n_1} \left(1 - \frac{\gamma}{s\mu + l\gamma} \right) = \frac{\gamma}{s\mu + n_1\gamma}.$$

Knowing that our customer will abandon at position j, the time to abandon, say $Z_{n_1}(j)$, is the sum of $n_1 - j + 1$ independent exponential random variables with parameters $s\mu + n_1\gamma, s\mu + (n_1 - 1)\gamma, \ldots, s\mu + j\gamma$. Averaging over all possibilities leads to

$$\mathbb{E}Z_{n_1}^k = \frac{\gamma}{s\mu + n_1\gamma} \sum_{j=1}^{n_1} \mathbb{E}Z_{n_1}^k(j).$$

For example, the expected value of Z_{n_1} may simply be written as

$$\mathbb{E}Z_{n_1} = \frac{1}{s\mu + n_1\gamma} \sum_{j=1}^{n_1} \frac{j\gamma}{s\mu + j\gamma}.$$

Using the results above combined with Equations (6.1)–(6.3), we obtain all moments of the random variables W_1, $W_{1,d}$ and $W_{1,d,s}$.

Low-Priority Customers

Our approach to derive the performance measures of type 2 customers is based on computing their virtual waiting time. Recall that the virtual waiting time is defined as the waiting time of an infinitely patient customer. For a new type 2 customer, we denote it by V_2. In what follows, we compute the k-th order moment of $W_{2,s}$ and $W_{2,r}$. Using the latter, all remaining performance measures are easily obtained thereafter.

Let us focus on the conditional waiting time of a type 2 customer given service, namely $W_{2,s}$. Recall that patience times are described by the random variable T. We have

$$F_{W_{2,s}}(t) = \frac{\mathbb{P}(V_2 < t, V_2 < T)}{\mathbb{P}(V_2 < T)}, \tag{6.9}$$

for $t \geq 0$. First, observe that $\mathbb{P}(V_2 < T) = P_{2,s}$. Second, $\mathbb{P}(V_2 < t, V_2 < T) = \int_0^t e^{-\gamma x} f_{V_2}(x)\mathrm{d}x$. A new type 2 arrival who finds at least one idle server with probability $1 - P_d$, will immediately enter service. If not, assume that she finds $n = n_1 + n_2$ waiting customers. Thus, we may write for $t \geq 0$

$$\mathbb{P}(V_2 < t, V_2 < T) = (1 - P_d) \cdot 1 + \int_0^t e^{-\gamma x} \sum_{n=0}^{\infty} p(n) f_{V_{2,n}}(x)\mathrm{d}x,$$

where $V_{2,n}$ is the conditional virtual waiting time of a new type 2 customer, given that upon arrival she finds in total n waiting customers in both queues. Her virtual waiting is not affected by all future type 2 arrivals because the discipline of service within queue 2 is FCFS. However, all future type 1 arrivals have to be considered because of their higher priority. Note also that this virtual waiting time does not depend on the couple (n_1, n_2) but on the total number of customers ahead of her $n = n_1 + n_2$ (common distribution of service and patience times for both

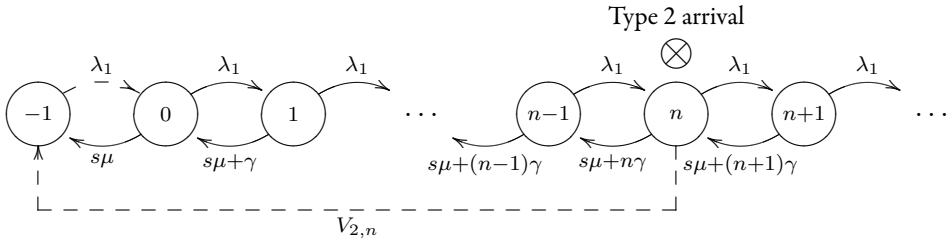


Figure 6.3. Virtual waiting time of a type 2 arrival finding n customers in queues 1 and 2, FCFS.

customer types). As a consequence $V_{2,n}$ can be seen as the first-passage time at state -1 starting at state n in the birth-death process as shown in Figure 6.3. It is the time to empty the queue ahead of our customer and in addition one server becomes idle to handle her.

By considering a single-class $M/M/s+M$ queue (with mean arrival rate λ_1), one may see that $V_{2,n}$ is equivalent to the duration of an n-busy period, for $n \geq 0$. Let us denote the latter by BP_{n,λ_1}, $V_{2,n} \equiv BP_{n,\lambda_1}$, for $n \geq 0$. Equation (6.9) then becomes

$$F_{W_{2,s}}(t) = \frac{1}{P_{2,s}} \left\{ 1 - P_d + \int_0^t e^{-\gamma x} \sum_{n=0}^{\infty} p(n) f_{BP_{n,\lambda_1}}(x) \mathrm{d}x \right\}, \tag{6.10}$$

for $t \geq 0$. Taking the derivative in t on both sides of Equation (6.10), we obtain

$$f_{W_{2,s}}(t) = \frac{1}{P_{2,s}} \sum_{n=0}^{\infty} p(n) e^{-\gamma t} f_{BP_{n,\lambda_1}}(t), \tag{6.11}$$

for $t \geq 0$. For the rest of the chapter, we denote by $\widetilde{F}_X(x)$, for $x \in \mathbb{R}^+$, the Laplace-Stieltjes transform of the pdf $f_X(\cdot)$ of a random variable X. Note that the Laplace-Stieltjes transform of $e^{-\gamma t} f_{BP_{n,\lambda_1}}(t)$ is $\widetilde{F}_{BP_{n,\lambda_1}}(x+\gamma)$, for $x \in \mathbb{R}^+$. Applying next the Laplace-Stieltjes transform to Equation (6.11) implies

$$\widetilde{F}_{W_{2,s}}(x) = \frac{1}{P_{2,s}} \sum_{n=0}^{\infty} p(n) \widetilde{F}_{BP_{n,\lambda_1}}(x+\gamma), \tag{6.12}$$

for $x \in \mathbb{R}^+$. Using Equation (6.12), one can obtain any k-th order moment of $W_{2,s}$, for $k \geq 1$. It is given by

$$(-1)^k \frac{\mathrm{d}^k \widetilde{F}_{W_2}(x)}{\mathrm{d}x^k}\bigg|_{x=0},$$

for $k \geq 1$. Thus

$$\mathbb{E}W_{2,s}^k = \frac{(-1)^k}{P_{2,s}} \sum_{n=0}^{\infty} p(n) \widetilde{F}^{(k)}_{BP_{n,\lambda_1}}(\gamma),$$

where $h^{(k)}(\cdot)$ denotes the k-th derivative of a function $h(\cdot)$, for $k \geq 1$.

Let us now focus on the conditional waiting time of a type 2 customer given abandonment, $W_{2,r}$. We have

$$F_{W_{2,r}}(t) = \frac{\mathbb{P}(T < t, V_2 > T)}{\mathbb{P}(V_2 > T)},$$

for $t \geq 0$. First, observe that $\mathbb{P}(V_2 > T) = P_{2,r}$. Second, we may write

$$\mathbb{P}(T < t, V_2 > T) = \int_0^t \gamma e^{-\gamma x}(1 - F_{V_2}(x))\mathrm{d}x,$$

for $t \geq 0$. As a consequence, we obtain after some algebra

$$F_{W_{2,r}}(t) = \frac{1}{P_{2,r}} \left\{ 1 - e^{-\gamma t} - \int_0^t \gamma e^{-\gamma x} \left(1 - P_d + \sum_{n=0}^{\infty} p(n) F_{BP_{n,\lambda_1}}(x) \right) \mathrm{d}x \right\}, \tag{6.13}$$

for $t \geq 0$. Taking the derivative in t on both sides of Equation (6.13) leads to

$$f_{W_{2,r}}(t) = \frac{\gamma}{P_{2,r}} \left(P_d e^{-\gamma t} - e^{-\gamma t} \sum_{n=0}^{\infty} p(n) F_{BP_{n,\lambda_1}}(t) \right), \tag{6.14}$$

for $t \geq 0$. Using that the Laplace-Stieltjes transform of $F_{BP_{n,\lambda_1}}(t)$ is $\frac{1}{x}\widetilde{F}_{BP_{n,\lambda_1}}(x)$, for $x \in \mathbb{R}^+$, and applying the Laplace-Stieltjes transform to Equation (6.14) implies

$$\widetilde{F}_{W_{2,r}}(x) = \frac{\gamma}{P_{2,r}(x+\gamma)} \left(P_d - \sum_{n=0}^{\infty} p(n) \widetilde{F}_{BP_{n,\lambda_1}}(x+\gamma) \right),$$

for $x \in \mathbb{R}^+$. This finishes the characterization of $W_{2,s}$ and $W_{2,r}$. One can now use Equations (6.1)–(6.3) to obtain all moments of the remaining random variables W_2, $W_{2,d}$ and $W_{2,d,s}$.

To close the discussion, we note that one can obtain the expected queue lengths Q_1 and Q_2 (given by Equation (6.7)) by using the expressions for $\mathbb{E}W_1$ and $\mathbb{E}W_2$ derived in this section and by applying Little's law, $\lambda_m \mathbb{E}W_m = Q_m$, for $m \in \{1, 2\}$.

6.3.2 Analysis of Model$_{\text{LCFS}}$

Similarly to the previous subsection, we compute here for Model$_{\text{LCFS}}$ the k-th order moment of $W_{m,s}$ and $W_{m,r}$, which also allows to derive the k-th order moment of W_m, $W_{m,d}$ and $W_{m,d,s}$, for $k \geq 1$ and $m \in \{1, 2\}$. We use the same approach based on the computation of the virtual waiting time of high- and low-priority customers.

High-Priority Customers

Let us consider a new "tagged" type 1 arrival and assume that she is infinitely patient. We denote her virtual waiting time by V_1. If she finds at least one idle server with probability $1 - P_d$, she immediately enters service. So, her virtual waiting time is zero. In the complementary case (all servers are busy), she is queued. Type 2 customers already waiting and those who arrive later are ignored because of their lower priority. Also, because the discipline of service within queue 1 is LCFS, type 1 customers already waiting in the queue are ignored. Thus, the conditional virtual waiting time, given delay for a new type 1 arrival is independent of the state of the two queues. Let us denote this conditional virtual waiting time by $V_{1,d}$, $V_1 = P_d V_{1,d}$. One can see that $V_{1,d}$ is the first-passage time at state -1 starting at state 0 in the birth-death process as shown in Figure 6.4.

From Figure 6.4, one can see that $V_{1,d}$ is equivalent to the duration of a 0-busy period in an $M/M/s + M$ queue with mean arrival rate λ_1, denoted by BP_{0,λ_1}. In a similar way as in Subsection 6.3.1, we characterize $W_{1,s}$ as follows. We have

$$F_{W_{1,s}}(t) = \frac{\mathbb{P}(V_1 < t, V_1 < T)}{\mathbb{P}(V_1 < T)},$$

for $t \geq 0$. We then obtain after some algebra

$$F_{W_{1,s}}(t) = \frac{1}{P_{1,s}} \left\{ 1 - P_d + P_d \int_0^t e^{-\gamma x} f_{V_{1,d}}(x) \mathrm{d}x \right\}, \tag{6.15}$$

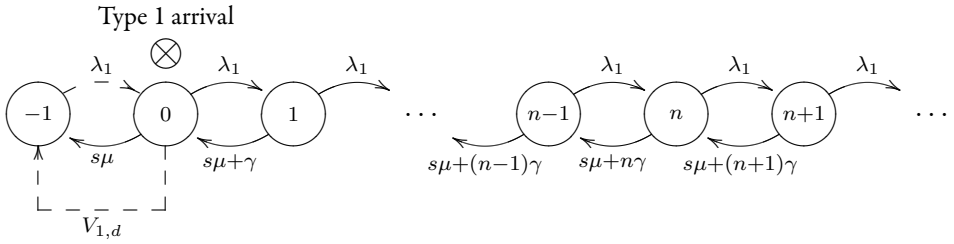


Figure 6.4. Virtual waiting time of a type 1 arrival, LCFS.

for $t \geq 0$. Using $V_{1,d} \equiv BP_{0,\lambda_1}$ and taking the derivative in t on both sides of Equation (6.15) gives

$$f_{W_{1,s}}(t) = \frac{P_d}{P_{1,s}} e^{-\gamma t} f_{BP_{0,\lambda_1}}(t),$$

for $t \geq 0$, which by applying the Laplace-Stieltjes transform leads to

$$\widetilde{F}_{W_{1,s}}(x) = \frac{P_d}{P_{1,s}} \widetilde{F}_{BP_{0,\lambda_1}}(x+\gamma),$$

for $x \in \mathbb{R}^+$. Finally, we obtain

$$\mathbb{E}W_{1,s}^k = (-1)^k \frac{P_d}{P_{1,s}} \widetilde{F}^{(k)}_{BP_{0,\lambda_1}}(\gamma),$$

for $k \geq 1$. We now move to characterize $W_{1,r}$. We have

$$F_{W_{1,r}}(t) = \frac{\mathbb{P}(T < t, V_1 > T)}{\mathbb{P}(V_1 > T)},$$

for $t \geq 0$, which implies after some simplifications

$$F_{W_{1,r}}(t) = \frac{1}{P_{1,r}} \left\{ 1 - e^{-\gamma t} - \int_0^t \gamma e^{-\gamma x} \left(1 - P_d + P_d F_{BP_{0,\lambda_1}}(x) \right) \mathrm{d}x \right\}, \tag{6.16}$$

for $t \geq 0$. Taking the derivative in t on both sides of Equation (6.16) leads to

$$f_{W_{1,r}}(t) = \frac{P_d \gamma}{P_{1,r}} \left(e^{-\gamma t} - e^{-\gamma t} F_{BP_{0,\lambda_1}}(t) \right), \tag{6.17}$$

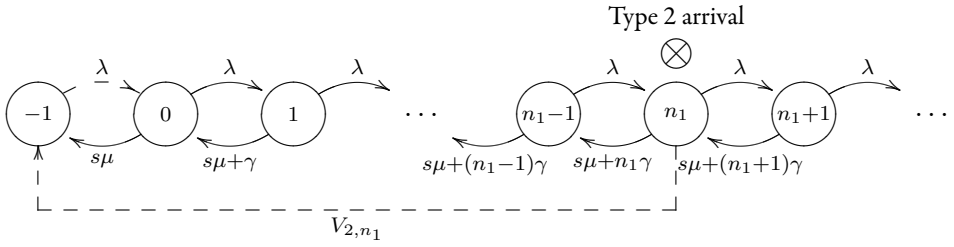


Figure 6.5. A new type 2 arrival arriving to an LCFS queue.

for $t \geq 0$. We now apply the Laplace-Stieltjes transform to Equation (6.17) and obtain

$$\widetilde{F}_{W_{1,r}}(x) = \frac{P_d \gamma}{P_{1,r}(x+\gamma)} \left(1 - \widetilde{F}_{BP_{0,\lambda_1}}(x+\gamma)\right),$$

for $x \in \mathbb{R}^+$. Finally, we close the discussion by mentioning that again one can use Equations (6.1)–(6.3) to obtain all moments of the remaining random variables W_1, $W_{1,d}$ and $W_{1,d,s}$.

Low-Priority Customers

Our approach again relies on determining the virtual waiting time of an infinitely patient type 2 customer. Consider such a customer. She will have a zero virtual waiting time with probability $1 - P_d$. With the complementary probability, she is queued. In the latter case, she will get priority over the type 2 customers already waiting. What matters for her virtual waiting time are the type 1 customers already waiting in queue 1 (denoted by n_1, $n_1 \geq 0$), as well as all future arrivals of types 1 and 2. Let us denote the conditional virtual waiting time of a type 2 customer, given a busy system and n_1 customers in queue 1, by V_{2,n_1}, $n_1 \geq 0$. One can see that V_{2,n_1} is the first-passage time at state -1 starting at state n_1 in the birth-death process as shown in Figure 6.5. It is then easy to see that V_{2,n_1} is equivalent to the duration of an n_1-busy period, say $BP_{n_1,\lambda}$, of an $M/M/s+M$ queue (with mean arrival rate $\lambda = \lambda_1 + \lambda_2$), $V_{2,n_1} \equiv BP_{n_1,\lambda}$.

Similarly to Equation (6.10), but by conditioning here on the state of queue 1,

we obtain

$$F_{W_{2,s}}(t) = \frac{1}{P_{2,s}} \left\{ 1 - P_d + \int_0^t e^{-\gamma x} \sum_{n_1=0}^{\infty} p_1(n_1) f_{BP_{n_1,\lambda}}(x) \mathrm{d}x \right\}, \tag{6.18}$$

for $t \geq 0$. Taking the derivative in t on both sides of Equation (6.18) gives

$$f_{W_{2,s}}(t) = \frac{1}{P_{2,s}} \sum_{n_1=0}^{\infty} p_1(n_1) e^{-\gamma t} f_{BP_{n_1,\lambda}}(t),$$

for $t \geq 0$. Next, we may write

$$\widetilde{F}_{W_{2,s}}(x) = \frac{1}{P_{2,s}} \sum_{n_1=0}^{\infty} p_1(n_1) \widetilde{F}_{BP_{n_1,\lambda}}(x+\gamma), \tag{6.19}$$

for $x \in \mathbb{R}^+$. In a similar way as that for the FCFS case, but by using the random variable $BP_{n_1,\lambda}$ and averaging over all queue 1 states, we have

$$F_{W_{2,r}}(t) = \frac{1}{P_{2,r}} \left\{ 1 - e^{-\gamma t} - \int_0^t \gamma e^{-\gamma x} \left(1 - P_d + \sum_{n_1=0}^{\infty} p_1(n_1) F_{BP_{n_1,\lambda}}(x) \right) \mathrm{d}x \right\},$$

for $t \geq 0$, and

$$\widetilde{F}_{W_{2,r}}(x) = \frac{\gamma}{P_{2,r}(x+\gamma)} \left(P_d - \sum_{n_1=0}^{\infty} p_1(n_1) \widetilde{F}_{BP_{n_1,\lambda}}(x+\gamma) \right), \tag{6.20}$$

for $x \in \mathbb{R}^+$. Again, one can use Equations (6.1)–(6.3) to obtain all moments of the remaining random variables W_2, $W_{2,d}$ and $W_{2,d,s}$.

Note that for all cases analyzed above (any customer type and any discipline of service), one can check the relation $\mathbb{E}W_m^k = P_{m,s}\mathbb{E}W_{m,s}^k + P_{m,r}\mathbb{E}W_{m,r}^k$, for $k \geq 1$ and $m \in \{1, 2\}$. In what follows, we do it for type 2 customers and $\text{Model}_{\text{LCFS}}$. It suffices to prove that $\widetilde{F}_{W_2}(x) = P_{2,s}\widetilde{F}_{W_{2,s}}(x) + P_{2,r}\widetilde{F}_{W_{2,r}}(x)$, for $x \in \mathbb{R}^+$. On the one hand, using Equations (6.19) and (6.20), we state that

$$P_{2,s}\widetilde{F}_{W_{2,s}}(x) + P_{2,r}\widetilde{F}_{W_{2,r}}(x) = \frac{\gamma P_d}{x+\gamma} + \frac{x}{x+\gamma} \sum_{n_1=0}^{\infty} p_1(n_1) \widetilde{F}_{BP_{n_1,\lambda}}(x+\gamma), \tag{6.21}$$

for $x \in \mathbb{R}^+$. On the other hand, we may write

$$F_{W_2}(t) = 1 - \mathbb{P}(\min\{V_2, T\} > t) = 1 - \mathbb{P}(V_2 > t)\mathbb{P}(T > t), \tag{6.22}$$

for $t \geq 0$. We also have

$$\begin{aligned} \mathbb{P}(V_2 > t) &= 1 - \left\{ (1 - P_d) \cdot 1 + \sum_{n_1=0}^{\infty} p_1(n_1)\mathbb{P}(V_{2,n_1} < t) \right\} \\ &= P_d - \sum_{n_1=0}^{\infty} p_1(n_1) F_{BP_{n_1,\lambda}}(t), \end{aligned} \tag{6.23}$$

for $t \geq 0$. Then, Equations (6.22) and (6.23) lead to

$$F_{W_2}(t) = 1 - P_d e^{-\gamma t} + e^{-\gamma t} \sum_{n_1=0}^{\infty} p_1(n_1) F_{BP_{n_1,\lambda}}(t),$$

for $t \geq 0$, which implies

$$f_{W_2}(t) = \gamma P_d e^{-\gamma t} + e^{-\gamma t} \sum_{n_1=0}^{\infty} p_1(n_1) f_{BP_{n_1,\lambda}}(t) - \gamma e^{-\gamma t} \sum_{n_1=0}^{\infty} p_1(n_1) F_{BP_{n_1,\lambda}}(t), \tag{6.24}$$

for $t \geq 0$. Finally, after some algebra, we deduce from Equation (6.24) that

$$\widetilde{F}_{W_2}(x) = \frac{\gamma P_d}{x + \gamma} + \frac{x}{x + \gamma} \sum_{n_1=0}^{\infty} p_1(n_1) \widetilde{F}_{BP_{n_1,\lambda}}(x + \gamma), \tag{6.25}$$

for $x \in \mathbb{R}^+$. By comparing Equations (6.21) and (6.25), we finish the proof.

6.3.3 More than Two Customer Types

The analysis in Subsections 6.3.1 and 6.3.2 can be extended to a model with more than two customer types, for both FCFS and LCFS cases. In what follows, we provide indications about the approach to use. For FCFS or LCFS, consider the extended $M/M/s + M$ queueing model with k customer types, for $k > 2$. Type m has nonpreemptive priority over type l, for $1 \leq m < l \leq k$. We assume that for all customer types, patience as well as service times are still statistically identical.

Let us now focus on the performance measures of a type m customer with mean arrival rate λ_m, for $1 \leq m \leq k$.

First, we need to compute the stationary probabilities to have all servers busy and i waiting customers in queues $1, 2, \ldots, m$, denoted by $p_{1\to m}(i)$, and those to have i waiting customers in all queues, denoted by $p(i)$, for $i \geq 0$ and $1 \leq m \leq k-1$. To compute these probabilities, it suffices to use the two-class analysis of Subsection 6.2.3 by transforming the k-class $M/M/s+M$ queue into a two-class one. We do so by aggregating the first m types into a one type with mean arrival rate $\sum_{j=1}^{m} \lambda_j$, and the rest of types into a second one with mean arrival rate $\sum_{j=m+1}^{k} \lambda_j$, for $1 \leq m \leq k-1$. This allows to compute P_d and also the expected number of customers in queues $1, 2, \ldots, m$, denoted by $Q_{1\to m}$, for $1 \leq m \leq k-1$, and that in all queues, denoted by $Q_{1\to k} = Q$. Thus, the expected length of queue m is $Q_m = Q_{1\to m} - Q_{1\to m-1}$, for $1 \leq m \leq k$. We then obtain $P_{m,r} = \frac{\gamma Q_m}{\lambda_m}$, and $P_{m,s} = 1 - P_{m,r}$, for $1 \leq m \leq k$. In what follows, we focus on characterizing the random variables $W_{m,s}$ and $W_{m,r}$, which allows also to characterize the remaining random variables W_m, $W_{m,d}$ and $W_{m,d,s}$, for $1 \leq m \leq k$. We use a similar approach as in the previous subsections, with some changes that we mention next. Each time, the approach consists on finding an equivalent two-class queue.

Consider the k-class model working under FCFS. For $m = 1$, we aggregate types $2, \ldots, k$ into one type. We then apply the same analysis as for high-priority customers in Subsection 6.3.1. For $2 \leq m \leq k$, we aggregate types $1, \ldots, m$ into one high-priority type, and types $m+1, \ldots, k$ into one low-priority type. We thereafter use the stationary probabilities $p_{1\to m}(i)$, and duration of the i-busy period of a single-class $M/M/s+M$ queue with mean arrival rate $\sum_{j=1}^{m-1} \lambda_j$, for $i \geq 0$.

Consider now the k-class model working under LCFS. For $m = 1$, what we need is P_d and the duration of the 0-busy period in a single-class $M/M/s+M$ queue with mean arrival rate λ_1. For $2 \leq m \leq k$, we in turn aggregate types $1, \ldots, m-1$ into one high-priority type, and types $m, \ldots, k$ into one low-priority type. We thereafter use the stationary probabilities $p_{1\to m-1}(i)$, and the duration of the i-busy period of a single-class $M/M/s+M$ queue with mean arrival rate $\sum_{j=1}^{m} \lambda_j$ (since for a new type m arrival, future type m arrivals have priority over her), for $i \geq 0$. This closes the discussion about the extension to a model with more than two customer types.

Remark 6.1. In what follows, we discuss the extension of the analysis to a mixed model similar to the basic one described in Subsection 6.2.1. The difference is

that we allow the discipline of service in one of the two queues to be different from that in the other queue. For example, type 1 customers are served under FCFS, while type 2 customers are served under LCFS (or the opposite case). The extension is very easy to do. All the expressions for the stationary probabilities in Section 6.2 hold for the mixed model. Consider a given type. If it is served under FCFS (LCFS), then it suffices to apply the same analysis as shown for that type in Subsection 6.3.1 (Subsection 6.3.2). This finishes the characterization of the mixed model.

6.3.4 Numerical Illustration

In this subsection, we give a numerical illustration of the analysis above. We compare the performance measures of $\text{Model}_{\text{FCFS}}$ and $\text{Model}_{\text{LCFS}}$. In our experiments, we choose $\mu = 1$. We vary the abandonment rate and the system size, with $\lambda_1 = \lambda_2 = s/2$.

First, we consider the conditional expected waiting times given service and given abandonment, for each customer type. The results are shown in Figure 6.6. For a single-class $M/M/s + M$ queue, Jouini (2012) proved that FCFS (LCFS) maximizes (minimizes) the conditional expected waiting time given service, and minimizes (maximizes) that given abandonment. It is easy to extend these results to any customer type in our multiple-type models here. In practice, for example in a call center, the manager would then prefer to use LCFS in order to improve the waiting time before service of a given type. This is unfair from a customer perspective. Figure 6.6 reveals that opposed to $\mathbb{E}W_{m,r}$, $\mathbb{E}W_{m,s}$ is not highly impacted by the policy in queue m, for $m \in \{1, 2\}$. Then, an appropriate decision for a manager is to use FCFS for each customer type. First, it allows to preserve fairness between customers of the same level. Second, it allows to achieve a good $\mathbb{E}W_{m,s}$ not far from the optimal one. Third, it is optimal in order to minimize the conditional waiting time given abandonment.

We go further by giving the standard deviations of queueing delays for both customer types in Table 6.1 for $\gamma = 0.5$. As one can see, Table 6.1 gives further arguments in favor of FCFS. Values of standard deviations are indeed lower for FCFS than those for LCFS, except for the single-server case for type 2.

As expected we see from Figure 6.6 that performance improves in the system size, due to pooling effects. Also, we see that performance in terms of queueing delays improves in the abandonment rate γ. As γ increases, patience times decrease,

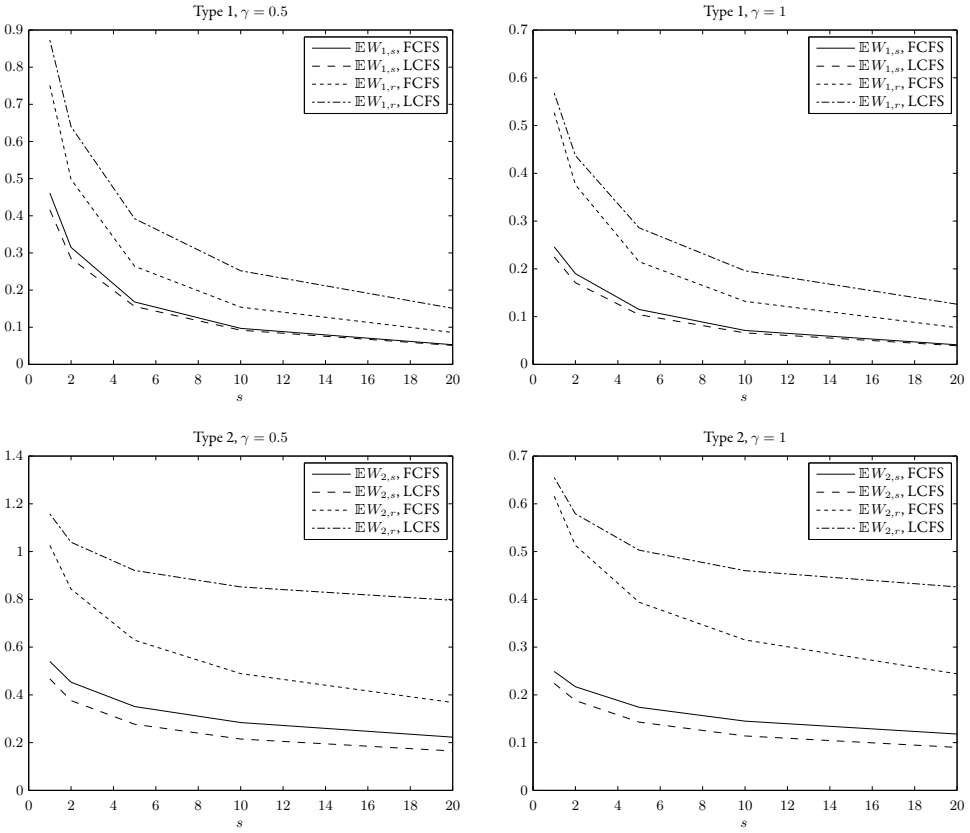


Figure 6.6. Conditional expected waiting times given service and given abandonment.

so fewer customers are present in the system, and as a consequence virtual delays improve. For each type, although the expected conditional waiting times given service and given abandonment (see Figure 6.6) do vary with the scheduling policy (FCFS, LCFS, etc.), the unconditional expected waiting times are as expected unchanged (see Table 6.1).

6.4 Conclusion

We considered multi-server nonpreemptive priority queueing systems in which customers wait for service for a limited time only and leave the system if service has not begun within that time. Practical examples of queueing systems with customer impatience include real-time telecommunication systems, inventory systems with

	Type 1, FCFS				Type 2, FCFS			
s	$\mathbb{E}W_1$	$\sigma(W_1)$	$\sigma(W_{1,s})$	$\sigma(W_{1,r})$	$\mathbb{E}W_2$	$\sigma(W_2)$	$\sigma(W_{2,s})$	$\sigma(W_{2,r})$
1	0.539	0.720	0.702	0.728	0.713	0.977	0.910	1.017
2	0.347	0.474	0.468	0.477	0.563	0.795	0.752	0.831
5	0.177	0.249	0.247	0.253	0.408	0.589	0.570	0.611
10	0.100	0.144	0.143	0.148	0.316	0.457	0.448	0.466
20	0.054	0.080	0.079	0.083	0.241	0.346	0.342	0.343
	Type 1, LCFS				Type 2, LCFS			
s	$\mathbb{E}W_1$	$\sigma(W_1)$	$\sigma(W_{1,s})$	$\sigma(W_{1,r})$	$\mathbb{E}W_2$	$\sigma(W_2)$	$\sigma(W_{2,s})$	$\sigma(W_{2,r})$
1	0.539	0.807	0.719	0.927	0.713	1.069	0.887	1.216
2	0.347	0.569	0.513	0.711	0.563	0.923	0.755	1.121
5	0.177	0.327	0.303	0.467	0.408	0.765	0.614	1.033
10	0.100	0.201	0.189	0.315	0.316	0.662	0.524	0.985
20	0.054	0.116	0.111	0.197	0.241	0.570	0.446	0.948

Table 6.1. Comparison between standard deviations of queueing delays.

perishable items, and more. We considered two models: one where the discipline of service within each class of customers is FCFS, and another one where it is LCFS. For each customer type, we explicitly derived the Laplace-Stieltjes transforms of the unconditional waiting time, the conditional waiting time given service, and the conditional waiting time given abandonment. Numerical inversion methods for Laplace-Stieltjes transforms can be then used in order to obtain the cdf values of these random variables at any point of time. Moreover, we described the approach to extend the analysis to more than two customer types. The analysis in this chapter holds also for a priority queue with mixed policies, i.e., FCFS for the first type and LCFS for the second one, and vice versa. Finally, we provided some numerical experiments in which we showed how FCFS would be preferred by a manager in practice.

There are various ways for future research. A challenging and interesting step is to extend our approach to the case of many customer types with different mean service and patience times. It is also interesting to consider general service-time distributions. Another useful extension would be to consider protocols with mixed priorities, i.e., both preemptive and nonpreemptive priorities.

Bibliography

J. Abate and W. Whitt. A unified framework for numerically inverting Laplace transforms. *INFORMS Journal on Computing*, 18(4):408–421, 2006.

M.S. Aguir. *Modèles Stochastiques pour l'Aide à la Décision dans les Centres d'Appels*. PhD thesis, Ecole Centrale Paris, 2004.

M.S. Aguir, F. Karaesmen, O.Z. Akşin, and F. Chauvet. The impact of retrials on call center performance. *OR Spectrum*, 26(3):353–376, 2004.

O.Z. Akşin, M. Armony, and V. Mehrotra. The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations Management*, 16(6):665–688, 2007.

S. Aldor-Noiman, P.D. Feigin, and A. Mandelbaum. Workload forecasting for a call center: Methodology and a case study. *Annals of Applied Statistics*, 3(4): 1403–1447, 2009.

E. Altman and A.A. Borovkov. On the stability of retrial queues. *Queueing Systems*, 26(3/4):343–363, 1997.

S. Asmussen. *Applied Probability and Queues*. Springer, 2nd edition, 2003.

S. Asmussen and P.W. Glynn. *Stochastic Simulation: Algorithms and Analysis*. Springer, 2007.

F. Avram, A.E. Kyprianou, and M.R. Pistorius. Exit problems for spectrally negative Lévy processes and applications to (Canadized) Russian options. *The Annals of Applied Probability*, 14(1):215–238, 2004.

A.N. Avramidis, A. Deslauriers, and P. L'Ecuyer. Modeling daily arrivals to a telephone call center. *Management Science*, 50(7):896–908, 2004.

F. Baccelli and G. Hebuterne. On queues with impatient customers. In *Performance '81*, pages 159–179. North-Holland, 1981.

J. Bard and H. Purnomo. Short-term nurse scheduling in response to daily fluctuations in supply and demand. *Health Care Management Science*, 8(4):315–324, 2005.

R.E. Barlow and L.C. Hunter. Reliability analysis of a one-unit system. *Operations Research*, 9(2):200–208, 1961.

O. Baron and J. Milner. Staffing to maximize profit for call centers with alternate service-level agreements. *Operations Research*, 57(3):685–700, 2009.

F. Baskett, K.M. Chandy, R.R. Muntz, and F.G. Palacios. Open, closed, and mixed networks of queues with different classes of customers. *Journal of the American Statistical Association*, 22(2):248–260, 1975.

R. Batta, O. Berman, and Q. Wang. Balancing staffing and switching costs in a service center with flexible servers. *European Journal of Operational Research*, 177(2):924–938, 2007.

R. Bekker, O.J. Boxma, and O. Kella. Queues with delays in two-state strategies and Lévy input. *Journal of Applied Probability*, 45(2):314–332, 2008.

O. Berman and R.C. Larson. A queueing control model for retail services having back room operations and cross-trained workers. *Computers & Operations Research*, 31(2):201–222, 2004.

O.J. Boxma and P.R. de Waal. Multiserver queues with impatient customers. In J. Labetoulle and J.W. Roberts, editors, *Proceedings of the 14th International Teletraffic Congress*, pages 743–756, 1994.

A. Brandt and M. Brandt. On the $M(n)/M(m)/s$ queue with impatient calls. *Performance Evaluation*, 35(1):1–18, 1999.

A. Brandt and M. Brandt. Asymptotic results and a Markovian approximation for the $M(n)/M(n)/s + GI$ system. *Queueing Systems*, 41(1/2):73–94, 2002.

A. Brandt and M. Brandt. On the two-class $M/M/1$ system under preemptive resume and impatience of the prioritized customers. *Queueing Systems*, 47(1/2): 147–168, 2004.

L. Brown, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. Multifactor Poisson and gamma-Poisson models for call center arrival times. Working paper, 2004.

L.D. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association*, 100(469):36–50, 2005.

G. Casella and R.L. Berger. *Statistical Inference*. Duxbury Press, 2nd edition, 2002.

B.D. Choi, B. Kim, and J. Chung. M/M/1 queue with impatient customers of higher priority. *Queueing Systems*, 38(1):49–66, 2001.

B. Cleveland and J. Mayben. *Call Center Management on Fast Forward: Succeeding in Today's Dynamic Inbound Environment*. Call Center Press, 1st edition, 1997.

R.B. Cooper. *Introduction to Queueing Theory*. North Holland, 2nd edition, 1981.

D.J. Daley and L.D. Servi. Idle and busy periods in stable $M/M/k$ queues. *Journal of Applied Probability*, 35(4):950–962, 1998.

A. Dassios and S. Wu. Semi-Markov model for excursions and occupation time of Markov processes. Working paper, 2011.

R.H. Davis. Waiting-time distribution of a multi-server, priority queuing system. *Operations Research*, 14(1):133–136, 1966.

A. Deslauriers, P. L'Ecuyer, J. Pichitlamken, A. Ingolfsson, and A.N. Avramidis. Markov chain models of a telephone call center with call blending. *Computers & Operations Research*, 34(6):1616–1645, 2007.

R.A. Doney and M. Yor. On a formula of Tákacs for Brownian motion with drift. *Journal of Applied Probability*, 35(2):272–280, 1998.

S. Drekic and D.A. Stanford. Threshold-based interventions to optimize performance in preemptive priority queues. *Queueing Systems*, 35(1/4):289–315, 2000.

F.F. Easton and J.C. Goodale. Schedule recovery: Unplanned absences in service operations. *Decision Sciences*, 36(3):459–488, 2005.

P.D. Feigin. Analysis of customer patience in a bank call center. Working paper, 2005.

Z. Feldman, A. Mandelbaum, W.A. Massey, and W. Whitt. Staffing of time-varying queues to achieve time-stable performance. *Management Science*, 54(2): 324–338, 2008.

N. Gans, G.M. Koole, and A. Mandelbaum. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5(2):79–141, 2003.

O. Garnett, A. Mandelbaum, and M. Reiman. Designing a call center with impatient customers. *Manufacturing & Service Operations Management*, 4(3): 208–227, 2002.

L.V. Green and P.J. Kolesar. The pointwise stationary approximation for queues with nonstationary arrivals. *Management Science*, 37(1):84–97, 1991.

L.V. Green, P.J. Kolesar, and J. Soares. Improving the SIPP approach for staffing service systems that have cyclic demands. *Operations Research*, 49(4):549–564, 2001.

L.V. Green, P.J. Kolesar, and J. Soares. An improved heuristic for staffing telephone call centers with limited operating hours. *Production and Operations Management*, 12(1):46–61, 2003.

L.V. Green, P.J. Kolesar, and W. Whitt. Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management*, 16(1):13–39, 2007.

S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29(3):567–588, 1981.

M. Harchol-Balter, T. Osogami, A. Scheller-Wolf, and A. Wierman. Multi-server queueing systems with multiple priority classes. *Queueing Systems*, 51(3/4): 331–360, 2005.

J. Harrison and A. Zeevi. A method for staffing large call centers based on stochastic fluid models. *Manufacturing & Service Operations Management*, 7(1):20–36, 2005.

J.E. Hosford. Measures of dependability. *Operations Research*, 8(1):53–64, 1960.

A. Ingolfsson. Modeling the $M(t)/M/s(t)$ queue with an exhaustive discipline. Working paper, 2005.

A. Ingolfsson, E. Akhmetshina, S. Budge, Y. Li, and X. Wu. A survey and experimental comparison of service-level-approximation methods for nonstationary $M(t)/M/s(t)$ queueing systems with exhaustive discipline. *INFORMS Journal on Computing*, 19(2):201–214, 2007.

F. Iravani and B. Balcıoğlu. Approximations for the $M/GI/N + GI$ type call center. *Queueing Systems*, 58(2):137–153, 2008a.

F. Iravani and B. Balcıoğlu. On priority queues with impatient customers. *Queueing Systems*, 58(4):239–260, 2008b.

T. Jiménez and G.M. Koole. Scaling and comparison of fluid limits of queues applied to call centers with time-varying parameters. *OR Spectrum*, 26(3): 413–422, 2004.

G. Jongbloed and G.M. Koole. Managing uncertainty in call centers using Poisson mixtures. *Applied Stochastic Models in Business and Industry*, 17(4):307–318, 2001.

O. Jouini. Analysis of a last come first served queueing system with customer abandonment. *Computers & Operations Research*, 2012. To appear.

O. Jouini and A. Roubos. On multiple priority multi-server queues with impatience. Submitted, 2011.

O. Jouini, O.Z. Akşin, and Y. Dallery. Call centers with delay information: Models and insights. *Manufacturing & Service Operations Management*, 13(4):534–548, 2011a.

O. Jouini, G.M. Koole, and A. Roubos. Performance indicators for call centers with impatience. Submitted, 2011b.

E.P.C. Kao and S.D. Wilson. Analysis of nonpreemptive priority queues with multiple servers and two priority classes. *European Journal of Operational Research*, 118(1):181–193, 1999.

E.L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

O. Kella and U. Yechiali. Waiting times in the non-preemptive priority $M/M/c$ queue. *Communications in Statistics. Stochastic Models*, 1(2):257–262, 1985.

L. Kleinrock. *Queueing Systems, Volume I: Theory*. John Wiley & Sons, 1976.

G.M. Koole and S.A. Pot. A note on profit maximization and monotonicity for inbound call centers. *Operations Research*, 59(5):1304–1308, 2011.

G.M. Koole, B.F. Nielsen, and T.B. Nielsen. First in line waiting times as a tool for analyzing queueing systems. *Operations Research*, 2012. To appear.

B.W. Kort. Models and methods for evaluating customer acceptance of telephone connections. In *GLOBECOM '83*, pages 706–714. IEEE, 1983.

A.E. Kyprianou. *Introductory Lectures on Fluctuations of Lévy Processes with Applications*. Springer, 2006.

S. Liao, G.M. Koole, C. van Delft, and O. Jouini. Staffing a call center with uncertain non-stationary arrival rate and flexibility. *OR Spectrum*, 2011. To appear.

H.W. Lilliefors. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62(318): 399–402, 1967.

D.V. Lindley. The theory of queues with a single server. *Mathematical Proceedings of the Cambridge Philosophical Society*, 48(2):277–289, 1952.

J.D.C. Little. A proof of the queuing formula: $L = \lambda W$. *Operations Research*, 9 (3):383–387, 1961.

A. Mandelbaum and S. Zeltyn. The impact of customers' patience on delay and abandonment: some empirically-driven experiments with the M/M/n+G queue. *OR Spectrum*, 26(3):377–411, 2004.

A. Mandelbaum and S. Zeltyn. Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. *Operations Research*, 57(5): 1189–1205, 2009.

M. Mandjes and M.J.G. van Uitert. Transient analysis of traffic generated by bursty sources, and its application to measurement-based admission control. *Telecommunication Systems*, 15(3/4):295–321, 2000.

F.J. Massey. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.

V. Mehrotra, O. Ozlük, and R. Saltzman. Intelligent procedures for intra-day updating of call center agent schedules. *Production and Operations Management*, 19(3):353–367, 2010.

L. Nguyen-Ngoc and M. Yor. Some martingales associated to reflected Lévy processes. In *Séminaire de Probabilités XXXVIII*, pages 42–69. Springer, 2005.

A. Pacheco. Some properties of the delay probability in $M/M/s/s+c$ systems. *Queueing Systems*, 15(1/4):309–324, 1994.

A. Pechtl. Distributions of occupation times of Brownian motion with drift. *Journal of Applied Mathematics & Decision Sciences*, 3(1):41–62, 1999.

E.J. Pinker and R.C. Larson. Optimizing the use of contingent labor when demand is uncertain. *European Journal of Operational Research*, 144(1):39–55, 2003.

J. Pitman and M. Yor. Hitting, occupation and inverse local times of one-dimensional diffusions: martingale and excursion approaches. *Bernoulli*, 9 (1):1–24, 2003.

S.V. Pustova. Investigation of call centers as retrial queuing systems. *Cybernetics and Systems Analysis*, 46(3):494–499, 2010.

M.L. Puterman. *Markov Decision Processes*. John Wiley & Sons, 1994.

S.S. Rao. Queuing with balking and reneging in M/G/1 systems. *Metrika*, 12(1): 173–188, 1967.

J. Riordan. *Stochastic Service Systems*. John Wiley & Sons, 1962.

T. Robbins. *Managing Service Capacity Under Uncertainty*. PhD thesis, The Pennsylvania State University, 2007.

A. Roubos, S. Bhulai, and G.M. Koole. Flexible staffing for call centers with nonstationary arrival rates. Submitted, 2011.

A. Roubos, R. Bekker, and S. Bhulai. Service-level distribution of multi-server queues. Submitted, 2012a.

A. Roubos, G.M. Koole, and R. Stolletz. Service-level variability of inbound call centers. *Manufacturing & Service Operations Management*, 2012b. To appear.

L. Rozenshmidt. On priority queues with impatient customers: Stationary and time-varying analysis. Master's thesis, Technion, Israel Institute of Technology, 2007.

G. Rubino and B. Sericola. Interval availability analysis using operational periods. *Performance Evaluation*, 14(3/4):257–272, 1992.

H. Shen and J.Z. Huang. Forecasting time series of inhomogeneous Poisson processes with application to call center workforce management. *Annals of Applied Statistics*, 2(2):601–623, 2008a.

H. Shen and J.Z. Huang. Interday forecasting and intraday updating of call center arrivals. *Manufacturing & Service Operations Management*, 10(3):391–410, 2008b.

A. Sleptchenko. Multi-class, multi-server queues with non-preemptive priorities. Technical report, EURANDOM, Eindhoven University of Technology, 2003.

A. Sleptchenko and M. van der Heijden. An exact solution for the state probabilities of the multi-class, multi-server queue with preemptive priorities. *Queueing Systems*, 50(1):81–107, 2005.

M.A.J. Smith, T. Aven, R. Dekker, and F.A. van der Duyn Schouten. A survey on the interval availability distribution of failure prone systems. In C. G. Soares, editor, *Advances in Safety and Reliability*, pages 1727–1737, 1997.

S.G. Steckley, S.G. Henderson, and V. Mehrotra. Service system planning in the presence of a random arrival rate. Working paper, 2004.

S.G. Steckley, S.G. Henderson, and V. Mehrotra. Forecast errors in service systems. *Probability in the Engineering and Informational Sciences*, 23(2):305–332, 2009.

R. Stolletz. *Performance Analysis and Optimization of Inbound Call Centers*. Springer, 2003.

R. Stolletz. Approximation of the non-stationary $M(t)/M(t)/c(t)$-queue using stationary queueing models: The stationary backlog-carryover approach. *European Journal of Operational Research*, 190(2):478–493, 2008.

L. Takács. On certain sojourn time problems in the theory of stochastic processes. *Acta Mathematica Hungarica*, 8(1/2):169–191, 1957.

T. Takine. The nonpreemptive priority $MAP/G/1$ queue. *Operations Research*, 47(6):917–927, 1999.

J.W. Taylor. A comparison of univariate time series methods for forecasting intraday arrivals at a call center. *Management Science*, 54(2):253–265, 2008.

T. Tezcan. Optimal control of distributed parallel server systems under the Halfin and Whitt regime. *Mathematics of Operations Research*, 33(1):51–90, 2008.

D.J. Thomas. Measuring item fill-rate performance in a finite horizon. *Manufacturing & Service Operations Management*, 7(1):74–80, 2005.

D. Wagner. Analysis of mean values of a multi-server model with non-preemptive priorities and non-renewal inputs. *Communications in Statistics. Stochastic Models*, 13(1):67–84, 1997.

A.R. Ward and P.W. Glynn. A diffusion approximation for a Markovian queue with reneging. *Queueing Systems*, 43(1/2):103–128, 2003.

J. Weinberg, L.D. Brown, and J.R. Stroud. Bayesian forecasting of an inhomogeneous Poisson process with applications to call center data. *Journal of the American Statistical Association*, 102(480):1186–1199, 2007.

W. Whitt. Improving service by informing customers about anticipated delays. *Management Science*, 45(2):192–207, 1999.

W. Whitt. Engineering solution of a basic call-center model. *Management Science*, 51(2):221–235, 2005.

W. Whitt. Staffing a call center with uncertain arrival rate and absenteeism. *Production and Operations Management*, 15(1):88–102, 2006.

J. Yoo. *Queueing Models for Staffing Service Operations*. PhD thesis, University of Maryland, 1996.

S. Zeltyn and A. Mandelbaum. Call centers with impatient customers: Many-server asymptotics of the M/M/n+G queue. *Queueing Systems*, 51(3/4):361–402, 2005.

S. Zeltyn, Z. Feldman, and S. Wasserkrug. Waiting and sojourn times in a multi-server queue with mixed priorities. *Queueing Systems*, 61(4):305–328, 2009.

Summary

Service-Level Variability and Impatience in Call Centers

A call center is defined as a group of telephone agents whose principal business is serving callers over the telephone. This thesis only considers the type of call center in which outside callers initiate the call to the call center. (The usual notation in the field of queueing theory is used where callers are called customers and agents are called servers.) The task of a call center manager is to keep costs as low as possible, which means keeping the number of servers as few as possible. On the other hand, a call center often has to obey a contract that prescribes that the service level has to be above a certain level. An example of this is that at least 80% of the customers should wait no longer than 20 seconds before receiving service.

A call center operates in a volatile environment with a high level of uncertainty. For example, it is impossible to know beforehand how many customers will call on a given day, when they will call, and how long a conversation will last. Under the Markovian assumptions, i.e., customers arrive according to a Poisson process and service times are assumed to be exponentially distributed, a call center can be modeled by a queueing system that is easy to analyze. For a given number of servers, it is possible to determine the service level, and therefore it is possible to find the optimal number of servers such that the service-level target will be met. However, there is a considerable problem that has not been recognized in the literature yet: the variability in the service level.

Most models only consider long-term performance, i.e., the service-level estimate will be met only in the long run. In practice however, service levels will be reported as an average over short intervals that typically range from several hours up to a day at most. In such short intervals, the service level is a random variable with a significant variability. The estimate provided by a queueing system is then only the expectation. By considering the whole distribution of the service level, improved decisions can be made regarding the optimal number of servers.

Chapters 2 and 4 concern the characterization of the distribution of the service level. In Chapter 2 a closed-form approximation for the variance is constructed. Based on this approximation, the normal probability distribution is used to characterize the form of the service-level distribution. This approximation is justified in the limit. The approximations turn out to be very accurate for shorter intervals starting from two or three hours. Furthermore, using the service-level distribution it is shown how better decisions can be made. Chapter 4 analyzes the distribution of the service level in an exact way by utilizing a double Laplace-Stieltjes transform. Instead of the customer-average service level, the time-average proportion of the time that the virtual waiting-time process is below the acceptable waiting time is considered. The time average converges to the customer-average service level due to a property of the Poisson process. The distribution of the service level is used in Chapter 3 to ensure that the service level at the end of the day is above the required target. In a call center a long-term planning has to be made concerning the number of scheduled servers, of which the level of uncertainty is high. At the start of the day and during the day more information becomes available. Using the idea of flexible servers, that can be added or removed, the number of servers is variable over the day and becomes a decision variable. The problem is modeled as a Markov decision process, that allows to find the optimal policy to meet the service-level target at minimal costs.

The second theme of this thesis is the impatience of customers. Customers that cannot directly get contact with a server have to wait in a virtual queue until a server becomes available. Customers have the property to be impatient: when the waiting time becomes too large, they hang up. This abandonment process influences the performance measures and therefore has to be taken into consideration in the models. An additional problem is that the service level is no longer unambiguously defined. Multiple definitions are used in practice.

Chapter 5 considers the problem of impatient customers. This chapter studies the different service-level definitions, including all those used in practice. Moreover, two new models are introduced based on data from call centers. Both models are shown to fit reality very well. The different service-level definitions are compared through numerical analysis. It is shown what the effect is on the required number of servers. Chapter 6 deals with a call center with impatience and multiple priorities. Customers are grouped by their priority and high-priority customers get priority over customers with a lower priority. It is shown how performance measures can be obtained for both types of customers under two kinds of service disciplines.

Samenvatting

Variabiliteit van het Service Level en Ongeduld in Call Centers

Een call center kenmerkt zich door een groep telefonisten die als voornaamste taak heeft het voeren van telefoongesprekken. Dit proefschrift beschouwt alleen het type call center waarbij bellers zelf naar het call center bellen. (De gebruikelijke notatie op het gebied van wachtrijtheorie wordt gehanteerd waar bellers klanten worden genoemd en telefonisten bedienden worden genoemd.) Een manager van een call center wil de kosten zo laag mogelijk houden, wat in feite erop neerkomt dat het aantal bedienden zo laag mogelijk moet zijn. Aan de andere kant moet een call center vaak aan een contract voldoen dat eist dat het serviceniveau boven een bepaalde grens moet zijn. Een voorbeeld hiervan is dat ten minste 80% van de klanten een wachttijd niet langer dan 20 seconden moet hebben.

Een call center opereert in een dynamische omgeving met een hoge mate van onzekerheid. Zo is bijvoorbeeld niet van tevoren te zeggen hoeveel klanten zullen bellen op een dag, wanneer ze bellen en hoe lang een gesprek duurt. Onder de Markovaannamen, waarbij klanten volgens een Poissonproces aankomen en bedieningsduren exponentieel verdeeld worden verondersteld, kan een call center gemodelleerd worden als een wachtrijmodel die eenvoudig te analyseren is. Voor een gegeven aantal bedienden is het mogelijk om hiermee het service level te bepalen, waarmee ook het optimale aantal bedienden gevonden kan worden zodat aan de eis van het service level wordt voldaan. Er is echter een groot probleem dat nog niet opgemerkt is in de literatuur: de variabiliteit van het service level.

De meeste modellen beschouwen alleen lange-termijn gemiddelden. Dit houdt in dat de schatting van het service level alleen op zeer lange termijn wordt bereikt. In de praktijk worden de service levels echter gerapporteerd als een gemiddelde over intervallen van een aantal uren tot maximaal een dag. Op zo'n korte termijn is het service level een stochastische variabele met een significante variabiliteit. De schatting van het service level die door een model wordt geleverd, is dan slechts

de verwachting. Door rekening te houden met de gehele verdeling van het service level kunnen er betere beslissingen worden genomen omtrent het optimale aantal bedienden.

Hoofdstukken 2 en 4 betreffen het karakteriseren van de verdeling van het service level. In Hoofdstuk 2 wordt een benadering in gesloten vorm bepaald voor de variantie. Gebaseerd op deze benadering wordt de normale kansverdeling gebruikt voor de vorm van de verdeling van het service level. Deze benadering is gerechtvaardigd in de limiet. Voor korte intervallen vanaf twee à drie uur blijken de benaderingen ook zeer nauwkeurig te zijn. Aan de hand van deze verdeling wordt laten zien hoe betere beslissingen genomen kunnen worden. In Hoofdstuk 4 wordt de verdeling van het service level exact geanalyseerd aan de hand van een dubbele Laplace-Stieltjes getransformeerde. In plaats van het klantgemiddelde wordt het tijdgemiddelde beschouwd van de fractie tijd dat het virtuele wachttijdproces onder de toegestane wachttijd is. Vanwege een eigenschap van het Poissonproces convergeert het tijdgemiddelde naar het klantgemiddelde. In Hoofstuk 3 wordt de verdeling van het service level gebruikt om het service level aan het einde van de dag boven het gewenste niveau te krijgen. In een call center wordt lang van tevoren, waarbij nog veel onzeker is, een planning gemaakt over het aantal in te zetten bedienden. Op de dag zelf en gedurende de dag is meer bekend. Door gebruik te maken van zogenoemde flexibele bedienden, die bij- of afgeschakeld kunnen worden, is het aantal bedienden variabel. Het probleem wordt gemodelleerd als een Markovbeslissingsproces, waarmee de optimale strategie wordt gevonden om tegen minimale kosten het service level te halen.

Het tweede onderwerp van dit proefschrift is het ongeduld van klanten. Klanten die bellen en niet direct een bediende aan de lijn kunnen krijgen, moeten wachten in een virtuele wachtrij totdat een bediende vrijkomt. Klanten hebben de eigenschap ongeduldig te zijn: indien de wachttijd te groot wordt, haakt men af. Dit afhaakproces heeft invloed op de prestatiematen en moet daarom meegenomen worden in de modellen. Een bijkomend probleem is dat het service level nu niet meer eenduidig gedefinieerd is. In de praktijk komen zelfs meerdere definities voor.

Hoofdstuk 5 beschouwt het probleem van ongeduldige klanten. In dit hoofdstuk worden de verschillende definities van het service level bestudeerd, waaronder al die in de praktijk gebruikt worden. Bovendien worden er twee nieuwe modellen geïntroduceerd gebaseerd op data afkomstig van call centers. Beide modellen komen zeer goed overeen met de werkelijkheid. Aan de hand van numerieke experimenten wordt laten zien wat het effect van de verschillende definities van het service

level is op het benodigde aantal bedienden. Hoofdstuk 6 beschouwt een call center met ongeduld en prioriteiten. Klanten worden gegroepeerd aan de hand van hun prioriteit en klanten met een hoge prioriteit krijgen voorrang boven klanten met een lagere prioriteit. Er wordt laten zien hoe de prestatiematen verkregen kunnen worden voor beide typen klanten onder twee soorten bedieningsdisciplines.